Reconsidering the Nature of Moral Reality: Goal Theory Unpacked and Evaluated

Thesis by

Nicholas Paul Clarke

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

The School of Advanced Study (University of London)

September 2019

Acknowledgements

First, I would like to thank my supervisor, Michael Lacewing, whose support and expert guidance throughout has been invaluable.

Richard Carrier's assistance in clarifying certain aspects of his theory has also been of enormous help to me.

In addition, I am indebted to my wife, Lisa, and to my daughter, Emily, who have been unfailingly supportive, and who have patiently tolerated my enforced absences from family life.

Finally, I must also acknowledge my parents, Bruce and Marion, who always encouraged my academic studies, and who would have been tremendously proud to witness the culmination of these.

Declaration of authorship

I, Nicholas Paul Clarke, declare that this thesis titled, ‘Reconsidering the Nature of Moral Reality: Goal Theory Unpacked and Evaluated’ and the work presented in it are my own. I confirm that:

- Where I have consulted the published work of others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- This thesis has not been submitted for a degree at any other University.

X ______________________________

Nicholas Paul Clarke

Abstract

When we survey the metaethical and normative landscape, we observe a problem, with each of the familiar theories facing serious and at least partially unresolved challenges, and all scoring poorly on certain widely accepted criteria for theoretical adequacy. This thesis argues that there is a novel theory, viz. Richard Carrier’s Goal Theory of morality, which fares much better in this regard, plausibly satisfying all of the applicable adequacy criteria, in addition to being resistant to the dominant challenges faced by accounts of its type.

I argue that if there is a true moral system, then it is probably Goal Theory. The theory’s first and second-order commitments are unpacked, and a number of important questions and objections answered. Three of these objections I deem substantial enough to warrant more detailed treatments: (1) that there are *categorical* normative reasons, contra Goal Theory’s Humean account; (2) that normative facts and properties are ‘just too different’ from natural facts and properties to be reducible or identical to them, contra Goal Theory’s naturalistic account of normativity; (3) and that ethical egoism succumbs to a number of internal and external criticisms, undermining Goal Theory’s egoist account. I argue that none of these objections defeats Goal Theory. With regard to theoretical adequacy, I argue that Goal Theory plausibly satisfies *all* of the criteria identified.

Having unpacked and evaluated Goal Theory, I conclude that it plausibly succeeds where existing theories do not. In so doing, it yields an outcome that is both highly surprising and resistant to easy refutation, making a substantial contribution to our quest to establish the true nature of moral reality and the true content of morality.

Table of Contents

Chapter 1

Introduction

What is the true nature of moral reality? What is the true content of morality? And how might we best make progress in the quest to establish these, answering questions about the very enterprise of normative ethics, and about what is right and wrong, good and bad, and what we morally ought to do?

The first two questions are central to the study of metaethics and normative ethics, respectively. In response to these questions (and others where ethics comes into contact with metaphysics, epistemology, philosophy of mind, and philosophy of language), we find a multitude of theories competing to deliver the correct answers. Each of these theories has its advocates amongst ethicists and metaethicists. And we find an implicit answer to the third question in their ongoing efforts to further develop and defend their theories of choice, in the hope that eventually all substantive objections will be overcome.

However, when we survey the contemporary metaethical and normative landscapes, we observe a problem, insofar as all of the familiar theories of the nature of moral reality or content of morality struggle to answer serious objections faced by the theoretical viewpoints to which they belong, and all seem to inherit the poor scores of their respective viewpoints on certain widely accepted criteria for theoretical adequacy.[1] In light of this, none of these theories appears to be especially epistemically

[1] For an articulation of some of these theoretical adequacy criteria, see the General Introduction in: Russ Shafer-Landau and Terence Cuneo, *Foundations of Ethics: An Anthology*, ed. by Russ Shafer-Landau and Terence Cuneo (Oxford: Blackwell Publishing Ltd, 2007). In this context, I interpret theoretical adequacy to be a measure of a theory’s overall epistemic probability, with this being composed of such

probable, and none is likely to convince anyone not antecedently inclined to accept it. Thus, I think we are currently presented with a set of *prima facie* inadequate theories from which to select.

To see what I mean, consider a few of the major theoretical viewpoints in metaethics. Firstly, at one end of the metaphysical spectrum we find moral error theory, which holds that there is no moral reality at all.[2] According to error theorists, moral judgements express beliefs and are truth-apt (a cognitivist semantic claim), but there are no objective moral facts (a denial of the realist metaphysical thesis). As such, if they endorse a correspondence theory of truth, then they are led to conclude that our moral judgements are uniformly and systematically false, because there are no moral facts in the world of the kind required to make them true. Such a view has the theoretic virtue of ontological parsimony, insofar as its ontology contains no moral facts and properties at all, so it can hardly be charged with multiplying entities beyond necessity. At the same time, it is not at all conservative (where conservatism is also generally held to be a theoretic virtue), since it fails to respect a particular pre-theoretical belief that is widely held and strongly resistant to alteration after reflection, viz. that our moral discourse represents reality (so that genocide *really is* morally wrong, and giving to charity *really is* morally good, for example). The error-theorist does not believe the opposite. Instead, they deny that any actions have such basic moral properties.

Expressivism (including emotivism, prescriptivism, norm-expressivism, assertoric non-descriptivism, and quasi-realism) concurs with error theory regarding the ultimate metaphysical status of morality, and so is similarly parsimonious, but it then disagrees about the nature of moral judgement — denying that such judgements can even *be* true or false. According to the expressivist, our moral judgements might appear

elements as its plausibility, parsimony, and explanatory scope and power. These components of epistemic probability will be discussed in section 6.2.

[2] E.g. J.L. Mackie, *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin, 1977).

on the surface to be assertions of moral propositions, but, in reality, are expressions of such things as emotions, sentiments of approval or disapproval, dispositions to form sentiments of approval or disapproval, or our acceptance of norms.[3] And, by their very nature, these have no substantial truth conditions. As such, expressivism appeals to those disinclined to accept either moral realism (with its supposedly extravagant metaphysical claims) or moral error theory (with its view that moral discourse is systematically mistaken).

However, expressivism has been charged with ‘interpreting a discourse in an eccentric manner simply to avoid philosophical difficulties’.[4] This ‘eccentric’ interpretation then leaves the expressivist struggling to explain why, if moral discourse is genuinely non-assertoric and non-truth-apt, it should appear otherwise. First-wave expressivists, such as A.J. Ayer, hardly attempted to explain this. However, second-wave expressivists, such as Blackburn and Gibbard, do try to explain how, if our moral discourse is primarily expressive, we can vindicate our standard practice of attributing truth to moral claims, and how it can have the logical and grammatical features of an assertoric discourse. Quasi-realism, for example, seeks to explain and justify the realistic surface features of our language about ethics (e.g. that it appears assertoric, and that moral claims can figure in embedded contexts such as logical syllogisms) in a way that is compatible with expressivism. Moreover, it attempts to make sense of the apparent truth-aptness of moral claims by appealing to deflationary accounts of truth — thereby vindicating our standard practice of attributing truth to moral claims, without committing to moral facts that make those judgements true.

[3] E.g. A.J. Ayer, *Language, Truth and Logic*, 2nd edn (London: Gollancz, 1946[1936]); Simon Blackburn, *Spreading the Word* (Oxford: Oxford University Press, 1984); Alan Gibbard, *Wise Choices, Apt Feelings* (Oxford: Clarendon Press, 1990).

[4] Richard Joyce, 'Error Theory', in *International Encyclopedia of Ethics*, ed. by H. LaFollette (Hoboken, NJ: Wiley-Blackwell, 2013). Joyce was referring to non-cognitivism in general.

However, as Richard Joyce says, expressivism has trouble accounting for the authority of morality (because, if moral judgements are nothing more than an expression of agents' feelings, then why should anyone else not antecedently inclined to care about these feelings pay any attention?).[5] Moreover, expressivism also faces the Frege-Geach problem (also known as the embedding problem), whereby it seems to commit us to the view that the meaning of moral claims varies across asserted and unasserted contexts, when, in our ordinary moral discourse, we do not think that the meaning of moral claims varies in this fashion.[6] Expressivists have responded to the challenge (e.g. by developing a logic of attitudes, exploiting a deflationary or minimalist account of truth, and using descriptive content in a hybrid theory).[7] However, these responses have themselves been challenged by cognitivists, and it is unresolved whether any version of expressivism survives this challenge.[8]

Other problems facing expressivism include the problem of mind-dependence (whereby, if rightness and wrongness depend upon our sentiments, then it seems to be entailed that if our sentiments were to change or disappear, then rightness and wrongness would thereby change or disappear too); and the problem of the schizoid attitude (whereby it becomes difficult to take rightness and wrongness seriously if they are merely projections of our attitudes and sentiments — with expressivists in some sense holding moral commitments but also holding that they are ungrounded).

In contrast to these and other anti-realist views, realist accounts of morality (according to which there really are moral facts and properties, and things are morally

[5] Ibid.

[6] See: P. Geach, 'Assertion', *Philosophical Review,* (1964), 449-65.

[7] On the first response, see, for example: Blackburn, *Spreading the Word*. On the second: P. Horwich, 'Gibbard's Theory of Norms', *Philosophy and Public Affairs,* (1993), 67–79; D. Stoljar, 'Emotivism and Truth Conditions', *Philosophical Studies,* (1993), 81–101. On the last: S. Barker, 'Is Value Content a Component of Conventional Implicature?', *Analysis,* (2000), 268–79; F. Jackson, 'A Problem for Expressivism', ibid. (1998), 239–51; M. Ridge, 'Ecumenical Expressivism: Finessing Frege', *Ethics,* 116 (2006), 302–36.

[8] Challenges include: M. Schroeder, 'Hybrid Expressivism: Virtues and Vices', ibid.119 (2009), 257–309; M. van Roojen, 'Expressivism and Irrationality', *Philosophical Review,* (1996), 311-55; M. van Roojen, 'Expressivism, Supervenience and Logic', *Ratio,* (2005), 190-205; N. Zangwill, 'Moral Modus Ponens', ibid. (1992), 177-93.

right or wrong, good or bad, independent of our opinion or attitude towards them) face other challenges. For example, non-naturalist realism claims that moral facts and properties are irreducible and *sui generis* non-natural facts and properties; and, as such, they are not part of the subject matter of the natural and social sciences, or not based in empirical observation and induction, for example.[9] However, in positing new and unproven kinds of entities, such accounts incur a loss of ontological parsimony relative to accounts not positing these entities (and positing entities that do not fit within a naturalistic worldview, and thus do not build upon established precedents and known facts, may also bear negatively upon non-naturalism's plausibility). Thus, *ceteris paribus*, they would seem to be less epistemically probable. Moreover, such non-naturalist theories struggle to explain how *sui generis* non-natural moral facts and properties could supervene on natural facts and properties (such that any change in what an agent ought morally to do requires a change in the natural facts and properties of the case), to offer an adequate explanation for how we can come to know these moral facts and properties, and explain why we should be motivated by and have excellent reasons to comply with them.

Naturalist realism, by virtue of locating the domain of morality within the familiar natural world, may possess an adequate epistemology, insofar as its moral facts and properties are in principle discoverable by the methods of science. However, in identifying moral entities with natural ones, it then faces the difficulty of accounting for the *normativity* of moral facts and properties — where it is commonly held that normative facts and properties, as facts and properties concerned with what we have reason to or ought to do, seem to be *just too different* from natural ones to be reducible or identical to them. Moreover, it must answer semantic challenges, such as G.E.

[9] E.g. John McDowell, *Mind, Value, and Reality* (Cambridge, MA: Harvard University Press, 1998); R. Shafer-Landau, *Moral Realism: A Defense* (Oxford: Oxford University Press, 2003).

Moore’s Open Question argument (OQA), according to which ‘good’ as a term is indefinable and ‘good’ as a property is irreducible.[10]

Those naturalist accounts that are reductive may plausibly account for supervenience (with the moral world being identical to a subset of the non-moral world) and might score better in terms of ontological parsimony (in positing only already established natural entities), but they must then contend with the problem of multiple realizability.[11] By contrast, non-reductive varieties suffer from the loss of ontological parsimony incurred by positing new and unproven entities (in the form of [*sui generis*] irreducible natural moral facts and properties), and must face the reappearance of the supervenience issue (in being required to explain how its moral natural facts and properties supervene on, without being reducible to, distinct non-moral natural facts and properties).[12]

Similarly, extant theories of the *content* of morality struggle to answer serious objections faced by the theoretical viewpoints to which they belong, in addition to having difficulty meeting certain applicable adequacy criteria.

1.1 Aim and scope

In light of the above-mentioned problems facing existing theories, how might we best make progress in the quest to answer the central questions in metaethics and ethics? Perhaps efforts to further develop and refine some familiar theory (or variant of this)

[10] G.E. Moore, *Principia Ethica*, revised edn (Cambridge: Cambridge University Press, 1993[1903]).

[11] Examples of reductive naturalist accounts include Frank Jackson’s analytic moral functionalism and Peter Railton’s synthetic account: Frank Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press, 1998); Peter Railton, 'Moral Realism', *Philosophical Review*, 95 (1986), 163-207.

[12] Prominent non-reductive naturalist accounts include: Richard Boyd, 'How to Be a Moral Realist', in *Essays on Moral Realism*, ed. by G. Sayre-McCord (Ithaca, NY: Cornell University Press, 1988), pp. 181-228; David Brink, *Moral Realism and the Foundation of Ethics* (Cambridge: Cambridge University Press, 1989).

will eventually bear fruit, with all substantive objections being overcome. But what if there already exists some novel theory that resists the serious objections faced by the theoretical viewpoint to which it belongs, and that meets all of the widely accepted criteria for theoretical adequacy? Might there be such a novel but adequate theory, or is it merely a unicorn — something that we can conceive of, but which we are unlikely to ever find in the real world? In fact, I think that there is a plausible candidate for just such a theory, and my aim in this thesis will be to describe and defend it.

In order to make the case for this novel but *prima facie* adequate theory, I shall adopt the following methodology:

- Unpack the novel theory's commitments along various dimensions of ethics and metaethics.
- Advance positive arguments for the theory and its commitments.
- Critically evaluate and respond to dominant objections to these commitments.
- Assess the theory for adequacy against some applicable set of criteria.

The aim would not be to add one more partially inadequate theory to the landscape, but instead to adduce a theory that substantively *improves* upon existing ones, by repairing defects in these theories or suffering from fewer or less severe vulnerabilities.

In what follows, I shall offer a conciliatory invitation to be open-minded about a novel theory that is in many ways counter-intuitive, but which I think offers some real hope in the quests to determine the true nature of moral reality and the true content of morality. I expect this theory to meet with some initial scepticism. However, in submitting it for serious consideration, I hope to overcome this scepticism, opening the discussion with a view to making some significant progress on the matters at hand.

Throughout, I emphasise clarity and rigour, so that engaging with my arguments does not entail an exhaustive exercise in textual interpretation.

Taking the second-order features first, what characteristics should an adequate theory of the nature of moral reality possess? With reference to the theoretical adequacy criteria alluded to earlier, I submit that:

- It should plausibly account for the supervenience of the moral world upon the non-moral one, such that it is impossible for the former to differ unless there is also a difference in the latter. (Moral non-naturalism in particular struggles with this criterion, since it is unclear how it can account for *sui generis* non-natural moral facts and properties supervening upon non-moral ones.[13])
- It should have an adequate moral epistemology, accounting for how we can apprehend anything to be known within morality. (Moral non-naturalism again struggles here, insofar as it postulates moral facts and properties that cannot be apprehended by scientific investigation, and so must posit some special faculty or other means by which we can come to know them.)
- It should be able to account for the relatively greater depth and breadth of moral disagreement as compared with other areas of supposed objective truth, where a failure to do this is argued to undercut a theory's claim to *objective* moral judgements. (Objective moral realism is threatened here, because if morality were objective then, by all accounts, we should expect to see far less moral disagreement than we do.)
- It should have a semantics of moral discourse, supplying plausible answers to well-known semantic puzzles (e.g. Moore's Open Question Argument, and the

[13] E.g. Tristram McPherson, 'Ethical Non-Naturalism and the Metaphysics of Supervenience', in *Oxford Studies in Metaethics*, ed. by R. Shafer-Landau (Oxford: Oxford University Press, 2012), pp. 205-34.

Moral Twin Earth thought experiment[14]). (This is a challenge for moral naturalism, which attempts to define moral properties in natural terms.)

- It should be ontologically parsimonious, in not multiplying entities beyond necessity. (Moral non-naturalism and non-reductive ethical naturalism face difficulties with this criterion, in positing unproven entities — either in the form of *sui generis* non-natural moral facts and properties or [*sui generis*] irreducible natural ones.)
- It should be conservative, in preserving many of our existing moral beliefs that are widely held, supportive of other beliefs, and resistant to alteration after reflection. (Anti-realist views are particularly vulnerable here, since they deny that there are any objective moral facts, with moral judgements being either systematically and uniformly false [error theory], or else not in the business of representing reality at all [expressivism].)
- It should explain why it is that, necessarily, anyone who sincerely holds a moral view is motivated to some extent to comply with it. (Moral realism struggles here, because, on realism, moral judgements express beliefs, which do not seem to be intrinsically motivating.)
- Finally, it should explain how moral requirements entail excellent reasons for compliance. (Something that instrumentalist theories have difficulty explaining, since the reason-giving power of moral requirements is then contingent upon them serving one’s commitments.)

Different metaethicists would assign greater or lesser significance to each of these theoretical adequacy criteria — perhaps denying some altogether. Nonetheless, any second-order theory that plausibly satisfied them all, in addition to offering cogent

[14] See: Terry Horgan and Mark Timmons, 'New Wave Moral Realism Meets Moral Twin Earth', *Philosophy and Phenomenological Research,* 16 (1991), 447-65.

answers to the dominant objections faced by the theoretical viewpoint to which it belongs, would surely be a credible one, worthy of serious consideration by other philosophers.

In terms of first-order features, we might say that an adequate theory should at least explain what is right and wrong (giving us a clear way of getting answers to our questions about actual moral situations); be comprehensive (giving us answers, or at least a way of establishing such answers, that we can imagine applying to any situation); be consistent (not yielding conflicting results in different circumstances — perhaps by starting from some basic principles, and then applying these principles systematically to particular situations to get answers); defuse or explain possible conflicts between self-interest and morality; and explain why we should be moral. Moreover, it should also plausibly resist the dominant objections aimed at its commitments.

The primary aim of this thesis will be to see if my candidate theory withstands appropriate critical scrutiny. Given that addressing all possible objections (as well as counter-responses to my arguments) is outside the scope of the thesis, I am not aiming here for cast-iron proof. Instead, I would like to establish that the theory *plausibly* meets the evaluative criteria that have been set out. The candidate theory I shall propose is Richard Carrier’s Goal Theory of morality, on which moral ‘rightness’ for an agent in such and such circumstances just is the property that actions have when they best serve the strongest desire that a fully rational and sufficiently informed version of the agent would have in those circumstances (and so on for other moral properties, such as ‘goodness’ or ‘wrongness’). The theory is unknown within the academic literature, having so far only received a relatively high-level treatment from

Carrier, in the form of two book chapters aimed at an educated lay audience.[15] As might be expected from this, it is significantly underspecified and will require much unpacking in order to define and subsequently test its commitments.

In this thesis, rather than evaluating Goal Theory primarily as a *first*-order moral theory (investigating whether it correctly identifies the conditions under which actions are morally right or wrong, or motives and intentions morally good or bad, for example), I shall focus primarily (though not exclusively) upon its *second*-order features. I do this because I think there is a sense in which questions about the true *nature* of morality (regarding its status, foundations, and scope, for example) precede those about the true *content* of morality — insofar as we may not be able to establish the latter until we have made sufficient progress in establishing the former. For example, if it turns out that there are *no* moral facts or properties, but our preferred theory of truth entails that moral claims are true just in case there are facts or properties that make these claims true, then any specific moral claim will be false. Alternatively, if there are moral facts and properties, but we have no reliable way (even in principle) to apprehend these facts and properties, then (on the same theory of truth) we have no reliable way ever to know if any specific moral claim is true or false. Moreover, even if there are moral facts and properties, and we can apprehend them, if these facts and properties would not motivate people who sincerely hold a moral view to act accordingly, or give them excellent reasons for compliance with moral requirements, then we might care little about the true content of morality.

Thus, I suggest that a critical evaluation of its second-order features is the right place to focus an initial appraisal of Goal Theory — even though second-order

[15] Richard Carrier, *Sense and Goodness without God: A Defense of Metaphysical Naturalism* (Bloomington, IN: AuthorHouse, 2005); Richard Carrier, 'Moral Facts Naturally Exist (and Science Could Find Them)', in *The End of Christianity,* ed. by John Loftus (Amherst, NY: Prometheus, 2011), pp. 333-64. As Carrier observes, the latter was peer reviewed by several philosophers. In these chapters, Goal Theory is contrasted with Christian morality, but I set that aside.

positions seem to underdetermine first-order ones, much as first-order positions underdetermine applied ethical positions (with John Stuart Mill, George Berkeley, R.M. Hare, and G.E. Moore all being utilitarians, despite them holding radically different metaethical views), and much actual metaethical analysis is abstracted away from particular moral judgements. Having said that, I will nonetheless devote a whole chapter to critically evaluating and responding to objections to what I take to be Goal Theory's most contentious *first*-order feature, viz. its ethical egoism. Moreover, I shall also explain why I think it meets the kind of first-order adequacy criteria identified earlier.

Given that Goal Theory is unknown within academic philosophy (and is therefore untested against suitable critical scrutiny), if, upon suitable investigation, the theory turns out to plausibly meet all of the aforementioned evaluative criteria, then I think this would constitute significant progress in our quest to determine the true nature of moral reality and the true content of morality. As such, I think that an extended articulation and defence of Goal Theory will form an original and significant contribution — simultaneously advancing existing debates in the academic literature, enabling the theory to be usefully developed and refined, and (through the honest and charitable engagement of other philosophers) raising the level of argumentation and analysis. Not least, I hope that my result will be both highly surprising and resistant to easy refutation.

1.2 Overview

The thesis is structured in three main sections:

1. **A comprehensive introduction to Goal Theory.** In chapter 2, I shall describe and evaluate Goal Theory in more detail, referencing the relevant academic literature as applicable. I shall begin by presenting a positive argument for Goal Theory (as a first-order theory), and then unpack and defend its second-order metaphysical, epistemological, psychological, and other commitments, explaining how I think it improves upon other theories. Next, I shall respond to some questions and objections, and then point the way to a more exhaustive treatment of some particularly dominant objections to its commitments — two metaethical and one normative.
2. **A thorough evaluation of the objections mentioned above.** This will form the central core of the thesis. The specific objections are as follows: (1) that there are *categorical* normative reasons, contra Goal Theory's Humean account; (2) that normative facts and properties are *just too different* from natural facts and properties to be reducible or identical to them, contra Goal Theory's naturalistic account of normativity; and (3) that ethical egoism succumbs to a number of internal and external criticisms, spelling serious trouble for Goal Theory's egoist account. Critical evaluations of these objections will be the subject of chapters 3, 4, and 5 (respectively).
3. **An evaluation of Goal Theory's performance against the remaining theoretical adequacy criteria.** Most of the theoretical adequacy criteria will already have been dealt with during the sections mentioned above, but the

remaining ones will be evaluated in chapter 6. I shall argue that Goal Theory plausibly meets these criteria too.

Having unpacked Goal Theory's first and second-order commitments, advanced positive arguments for the theory and its commitments, critically evaluated and responded to dominant objections to these commitments, and shown that Goal Theory plausibly satisfies widely-accepted criteria for theoretical adequacy, I shall conclude by submitting that Goal Theory is a plausible candidate for the novel but adequate theory I described earlier. I shall then discuss some of the implications of this result and potential opportunities for future research. At the very least, I hope that other philosophers will accept my conciliatory invitation to be open-minded about Goal Theory, to treat it as seriously as they do other theories, and to contribute to the discussion that I have initiated. Even if Goal Theory ultimately succumbs to some other objection, I think that this process of engagement and development will still be a useful one for philosophy.

1.3 Basic definitions

Before I introduce Goal Theory, some definitions are in order. Firstly, in this thesis, I shall generally treat 'moral' and 'ethical' as being synonymous. I shall understand a *moral fact* to be a fact that consists of something's having a moral property, where I shall understand a *moral property* to be some value or evaluative property, such as rightness or goodness. *Moral realism* I understand as a metaethical view committed to the objectivity of ethics, on which there are moral facts independent of actual people's beliefs or attitudes (but not independent in the stronger sense of being mind-independent, such that there could be moral facts even if there were no people, other than as facts about hypothetical people, and not as facts independent of them).[16] *Moral anti-realism* is then the view that there are no such moral facts. Moral realism entails *cognitivism* (though not the opposite, since error theorists are cognitivists), which I shall understand as the position that a moral judgement expresses an agent's beliefs and that such judgements are therefore truth-apt (since beliefs have substantial truth conditions). *Non-cognitivism* is then the position that moral judgements express non-cognitive states (such as desires, emotions, or dispositions towards approval or disapproval), which have no substantial truth conditions ('substantial', to allow for minimalist accounts of truth-aptness). By a *representationalist* view, I shall mean one on which moral sentences represent a way reality could be. I shall understand a moral theory to be a *naturalist* one if judgements on that theory are rendered true or false by natural states of affairs (where, by a 'natural state of affairs', I shall mean a state of affairs that consists in the instantiation of a natural fact of the sort that can be employed by or referred to in the natural or social sciences).[17]

In terms of motivation, I shall understand *internalism about motivational judgement* as the position that there is a necessary, internal connection between making a sincere moral judgement and being motivated to act in accordance with that judgement, even if this motivation might be overridden by a greater motivation to act otherwise (where the context is clear, I shall henceforth refer to this merely as 'internalism'). *Externalism* is then the position that there is no such necessary connection, and any link between judgement and motivation is contingent upon an

[16] In the words of Michael Smith: 'Moral questions have correct answers, ... made correct by objective moral facts.' Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994), pp. 9, 13. Some may wish to distinguish my view, which they would call 'robust realism', from 'minimal realism', whereby our moral discourse is assertoric and sometimes true. By further contrast, what we might call normative realism (or metanormative realism) deals not merely with the set of *moral* facts, but with the set of *normative* facts, where I shall later argue that the latter constitutes a superset of the former.

[17] Moore's definition was the following: 'By nature then I do mean and have meant that which is the subject matter of the natural sciences, and also of psychology.' Moore, *Principia Ethica*, p. 92. I shall discuss the natural/non-natural distinction in section 2.2, where I shall attenuate my definition above.

agent's psychological state. The *Humean Theory of motivation* is the view that motivation requires the presence of a belief and an appropriately related and independently intelligible desire, where the desire takes the dominant role.[18] We might take the conjunction of moral realism, cognitivism, internalism, and Humean Psychology to represent the 'commonsense' view of morality, insofar as it tends to align with our everyday moral discourse. In contrast to the kind of internalism described above, I shall understand *internalism about reasons* to be the view that reasons for action must be internal in the sense that they are grounded in motivational facts about the agent, e.g. their desires or goals.

With regard to normativity, categoricity and the like, I shall understand a *normative* statement to be a statement to the effect that something *ought* to be done. A proposition, *p*, being normative can then be understood in two distinct senses: (1) that *p* is a proposition with ought-character (with no implication as to whether an agent ought to comply with *p*); and (2) as a statement about how one ought to act (the distinction between these two senses is analogous to the distinction between a *factual* proposition and a *true* proposition). A *reason* I shall understand to be a consideration that favours or opposes, that make something justified, appropriate, or legitimate; or the opposite. A *normative reason* I shall then understand as a reason an agent has for why they ought to act in a certain way. Unless I specify otherwise, whenever I refer to a reason I shall mean a *normative* reason, as opposed to a *motivating* (or explanatory) one (where this latter would be an answer to the question 'why did *A* do *x*?').

An *imperative* I shall understand as an ought-statement. A *hypothetical imperative* is then an ought-statement that states what an agent ought to do, given the

[18] David Hume, *A Treatise of Human Nature* (Oxford: Clarendon Press, 1888 [1968]). There is some debate as to whether Hume himself actually subscribed to this view e.g.: E. Millgram, 'Was Hume a Humean?', *Hume Studies*, 31 (1995), 75-93.

condition that a certain end is desired by the agent. By contrast, a *categorical imperative* states what an agent ought to do, independently of any such condition.

Chapter 2

The Goal Theory of morality

If you want your car to run well, you ought to change its oil with sufficient regularity. If you want to save the life of a patient on whom you are operating, you ought to sterilise your instruments. And if you want to build an enduring bridge, you ought not to employ brittle concrete.[1] These hypothetical imperatives are uncontroversial. But what if morality works the same way, with it being a system of hypothetical imperatives? Many doubt that hypothetical imperatives are sufficient to ground morality, but what if they are wrong about that? What might then be implied about the content and very enterprise of normative ethics?

In this chapter I shall describe, situate, and motivate Goal Theory — unpacking and defending its commitments along various dimensions of ethics and metaethics. One defining characteristic of Goal Theory is that it conceives of morality as just such a system of hypothetical imperatives, so in the course of the chapter I shall present one possible answer to the question of what might be implied about the content and very enterprise of normative ethics if morality is a system of hypothetical imperatives. I shall also respond to some objections and lay the foundations for the chapters to come. But let me begin with a positive argument for the theory.

[1] Examples taken from: Carrier, 'Moral Facts Naturally Exist (and Science Could Find Them)', p. 334.

2.1 A positive argument for Goal Theory

What do we mean by morality? There exist a multitude of normative moral theories, including utilitarianism, Kantianism, and Aristotelian virtue ethics, with each proposing its own conception of how we ought to live. However, one must distinguish here between these kinds of moral theories, and 'morality' — with the former providing explicit descriptions, explanations, and justifications of morality, but the latter being the actual *target* of this moral theorising. The latter shall be the focus of my attention here. We have both *descriptive* and *normative* senses of 'morality'. In the descriptive sense, we mean something like codes of conduct proposed by a group or society, or accepted by an individual for their own behaviour. By contrast, as a plausible basic schema for definitions of morality in the normative sense, then, based upon an a posteriori appeal to the (ethical theorist) community's linguistic intentions (i.e. what we find such people intend to refer to with their use of the word 'moral'), we might propose: a code of conduct that, given specified conditions, would be endorsed by all rational persons.[2] Here I am only concerned with the normative sense of 'morality', and will set aside the claim that descriptive moralities are the only moralities there are (such that there is *no* code of conduct that, given specified conditions, would be endorsed by all rational persons).[3]

Following Bernard Gert, we may improve considerably upon the aforementioned basic schema for definitions of 'morality' in the normative sense once we have at our disposal the notion of an 'informal public system'. By a 'public' system, we mean that all of those to whom it applies must understand it and that it

[2] See: Bernard Gert, *Morality: Its Nature and Justification, Revised Edition* (Oxford: Oxford University Press, 2005), pp. 10-13.

[3] Certain moral relativists, notably Jesse Prinz and David Wong, advance this claim: Jesse Prinz, *The Emotional Construction of Morals* (Oxford: Clarendon Press, 2007); David Wong, *Natural Moralities: A Defense of Pluralistic Relativism* (Oxford: Oxford University Press, 2006).

must not be irrational for them to use it in deciding what to do and in judging others to whom the system applies; and by ‘informal’ we mean that there is no decision procedure or authority that can settle all its controversial questions. Refining and formalising the definition of ‘morality’ accordingly, I would propose the following improved definition:

> **(Morality)**: an informal public system of imperatives that applies to all rational persons, governing behaviour that affects others.[4]

This public system of imperatives will then generate rules, ideals, and virtues. Whether these are a function of what produces the best consequences (to the agent themselves, or all agents in aggregate), of what behaviours are *intrinsically* best (regardless of consequences), or of what behaviours are entailed by the best virtues of character, for example, is a matter for further enquiry.

Now, as Gert goes on to say, amongst those who think that ‘morality’ refers to an informal public system of imperatives that applies to all rational persons, governing behaviour that affects others, it is commonly held that morality should never be *overridden*. By this, it is meant that moral imperatives should not be violated for non-moral reasons — with the ‘should’ in question typically understood as meaning ‘rationally should’, thereby making moral requirements rational requirements. I shall endorse this common view (and will revisit it later in this section). Incorporating it into the previous definition yields the following:

[4] This accords with Gert’s definition here: Bernard Gert, 'Morality', in *The Cambridge Dictionary of Philosophy*, ed. by Robert Audi (Cambridge: Cambridge University Press, 2015). More can be found here: Gert, *Morality: Its Nature and Justification, Revised Edition*. Gert does not explicitly say in his definition what morality is an informal public system *of*. However, I infer it to be something in the region of rules or imperatives, and have settled upon the latter for simplicity. As Gert observes, some would dispute the claim that morality only governs the behaviour that affects *others*. However, its inclusion in the definition is common.

> **(Morality$_2$)**: an informal public system of imperatives that applies to all rational persons, governing behaviour that affects others, and which should never be overridden.

Given this, the following proposition then follows straightforwardly:

> **(S$_1$)**: if there is a true moral system, then its system of imperatives supersedes all other imperatives for rational persons.

Here I say ‘if there is a true moral system’, because I allow for the possibility that there is no actual moral system that meets the specified criterion (just as we might say that if there is a true *unicorn*, then it is a horse with a single straight horn projecting from its forehead — but discover upon investigation that there is nothing in the world that meets this criterion). Now, (S$_1$) may be rewritten in terms of ‘ought’ language as follows:

> **(S$_2$)**: if there is a true moral system, then its system of imperatives dictates what rational persons ought most to do.

However, I would now argue that there is an identity (of property) between statements of the form ‘one ought most to do *x*’ and statements of the form ‘when fully rational and sufficiently informed one *will do x*’[5], with them having the same

[5] By ‘fully rational’, I mean nothing more than making entirely deductively valid or inductively forceful inferences (i.e. without fallacy), and not falling prey to any cognitive biases or suchlike. By ‘sufficiently informed’ I do not imply omniscience, merely that a person in question possesses sufficient relevant non-moral information (e.g. of their desires and the consequences of their actions).

extension, and it therefore being possible to substitute one for the other without changing the reference (just as with, say, 'justified true belief' and 'knowledge', and with 'H_2O' and 'water').[6] To see why, imagine that, when fully rational and sufficiently informed, person *P will* do *x* in *C*. In that case, what *ought P* to do in *C*?[7] Either it is *x*, or it is ~*x*. Consider the latter. Any sense of 'ought' that recommended doing ~*x* would be proposing that persons *ought* to act in a way that they would only freely do if they were (somewhat) irrational or uninformed. However, that would surely be a bizarre and implausible connotation of 'ought' — irrelevant to our ideal conduct, since the more we take care to act rationally and informedly, the less likely we are, *ceteris paribus*, to do what such an 'ought' commands. In the limit, being fully rational and sufficiently informed, we would be guaranteed, *ex hypothesi*, to act otherwise. If fully rational and sufficiently informed persons would not act as they 'ought' to do, then it is hard to conceive of how such a sense of 'ought' might be rationally justified. Thus, I submit that if, when fully rational and sufficiently informed, person *P will* do *x* in *C*, then that *P ought* to do ~*x* in *C* would very probably be false — thereby making it very probably true that *P ought* to do *x*. By the same logic, it can be inferred that if it is *not* the case that, when fully rational and sufficiently informed, person *P* will do *x* in *C* (i.e. *P* will do ~*x* in *C*), then it is very probably *not* the case that they *ought* to do *x* in *C*.

Thus, I would argue that: (1) if it is the case that *A* (i.e. when fully rational and sufficiently informed, person *P* will do *x* in *C*), then it is (very probably) the case that *B* (i.e. *P* ought to do *x* in *C*); and (2) if it is the case that ~*A*, then it is (very

[6] These statements may not be *semantically* equivalent (even on a Fregean concept, where S_1 and S_2 would be semantically equivalent if they have the same truth value — since e.g. 'Mary believes she ought most to do *x*' and 'Mary believes that, when fully rational and sufficiently informed, she will do *x*' may have different truth values). However, I do not think semantic equivalence is required here, with property identity sufficing for my purpose.

[7] For convenience, I am henceforth omitting the qualifier 'most', but it should be assumed.

probably) the case that ~*B*. The former is a *necessary* condition for the kind of identity I propose between *A* and *B*. So, for example, if there is an identity between water and H_2O, then it is necessarily the case that if I am holding a cup of H_2O, then it is the case that I am holding a cup of water. However, this is not a *sufficient* condition for identity. After all, we might say, for example, that if it is the case that I am feeding a domestic cat, then it is the case that I am feeding a feline. As such, condition (1) is met. However, there can be felines that are not domestic cats, so the reverse need not be true. However, once we incorporate condition (2), then this possibility is ruled out (i.e. the following statement would be false: if it is the case that I am not feeding a domestic cat, then it is the case that I am not feeding a feline). Thus, I would argue that in meeting both conditions, there is very probably an appropriate identity between the *A* and *B* in question (i.e. '*P* ought most to do *x*' is very probably the same as 'when fully rational and sufficiently informed, *P* will do *x*'). Accordingly, from (S_2), we may infer the following:

> **(S_3)**: if there is a true moral system, then its system of imperatives states what persons *will do*, when fully rational and sufficiently informed.[8]

What does it mean to say that, when fully rational and sufficiently informed, *P will do x* in circumstances *C*? Consider the widely accepted action-based theory of desire, on which for a person to *desire* Φ is for the person to be disposed to take whatever actions they believe are likely to bring about Φ. Desires so understood may then be stronger or weaker, with the strength of the desire constituted by its causal power concerning the control of action. As such, for one desire to be stronger than another is, *ceteris paribus*, for one to be disposed to act upon it, rather than acting upon the

[8] I assume hereafter that the persons concerned are able to act *freely*, i.e. not acting under coercion.

other desire (where one believes that the desires are satisfiable by distinct actions and not jointly satisfiable). On this theory, the predictive statement '*P will do x* in circumstances *C*, when fully rational and sufficiently informed' is then equivalent to the statement '*P will do x* in circumstances *C*, when fully rational and sufficiently informed, because rational persons will always be most disposed to act upon their strongest desire, and, when fully rational and sufficiently informed, *P* desires the results of doing *x* in *C* more than the results of doing ~*x* in *C*'. Thus, granting the action-based theory of desire, we now have:

> **(S_4)**: if there is a true moral system, then its system of imperatives states what persons will do, when fully rational and sufficiently informed, with this being what they would desire most when fully rational and sufficiently informed.

Translating this back into the language of 'oughts', using the identity that I proposed in the derivation of (S_3), we obtain the following equivalent statement:

> **(S_5)**: if there is a true moral system, then its system of imperatives states what rational persons ought most to do, with this being what they would desire most when fully rational and sufficiently informed.

Syllogistically, the foregoing may be reconstructed thus:

Argument 1

P1)	Morality is probably an informal public system of imperatives that applies to all rational persons, governing behaviour that affects others, and which should never be overridden. [From the definition (Morality$_2$)]
P2)	If morality is an informal public system of imperatives that applies to all rational persons (*R*), governing behaviour that affects others, and which should never be overridden, then, if there is a true moral system (*T*), its system of imperatives (*S*) supersedes all other imperatives for *R*.
C1)	Therefore, probably if there is a *T*, its *S* supersedes all other imperatives for *R*.
P3)	If a system of imperatives supersedes all other imperatives for *R*, then it dictates what *R* ought most to do. [Semantically equivalent statement]
C2)	Therefore, probably if there is a *T*, then its *S* dictates what *R* ought most to do.
P4)	There is very probably an identity (of property) between the statement '*P* ought most to do *x*' and the statement 'when fully rational and sufficiently informed, *P* will do *x*'.
C3)	Therefore, probably if there is a *T*, then its *S* very probably states what *R* will do, when fully rational and sufficiently informed.
P5)	What rational persons will do, when fully rational and sufficiently informed, is probably what they would *desire most* to do, when fully rational and sufficiently informed. [On the action-based theory of desire]

C4)	Therefore, probably if there is a *T*, then its *S* probably states what *R* will do, when fully rational and sufficiently informed, with this probably being what they would desire most when fully rational and sufficiently informed.
C5)	Therefore, [using the identity in P4] probably if there is a *T*, then its *S* probably states what *R* ought most to do, with this probably being what they would desire most when fully rational and sufficiently informed.

As a corollary to Argument 1, if we reframe (S_5) in terms of a particular rational person, *P*, in such and such circumstances, we may say that if there is a true moral system, *T*, then, on *T*, a moral ought is defined as follows:

> **(Ought)**: if and only if, when fully rational and sufficiently informed, *P* would desire the results of doing *x* in moral circumstances *C* more than they would desire the results of doing ~*x* in *C*, then *P* morally *ought* to do *x*.

The above may be written in the following equivalent form:

> **(Ought$_2$)**: *P* morally *ought* to do *x* in moral circumstances *C* just in case doing *x* in *C* would best serve the strongest desire of a fully rational and sufficiently informed version of *P*.

From the foregoing, we may also derive two additional corollaries. Firstly, written in terms of putative moral facts, we may say that on *T*:

> **(M_F)**: it is a *moral fact* that *P* ought to do *x* in moral circumstances *C* if, when fully rational and sufficiently informed, *P* would desire the results of doing *x* in *C* more than they would desire the results of doing ~*x* in *C*.

Working from (S_4) instead of (S_5), we may alternatively say that on *T*:

> **(M_{F2})**: it is a *moral fact* that, when fully rational and sufficiently informed, *P* will do *x* in moral circumstances *C*, with this being what they would desire most.

I shall understand an action to be morally *right* for an agent in such and such circumstances just in case it is a moral fact that they ought to do that thing in those circumstances. Given this, I may define moral *rightness* on *T* thus:

> **(M_R)**: action *x* is morally *right* in moral circumstances *C* for *P* just in case it is a moral fact that *P* ought to do *x* in *C*.

Alternatively, noting from this that moral rightness for *P* in *C* just is the property that an action has when *P* acts as they ought (i.e. in accord with the moral fact that *P* ought to do *x* in *C*), and the action dictated by this moral fact is the one that best serves the strongest desire of a fully rational and sufficiently informed version of *P*, then I may equivalently define moral rightness on *T* as:

> **(M_{R2})**: moral *rightness* for *P* in moral circumstances *C* just is the property that an action has when it best serves *P*'s true strongest desire in those circumstances.[9]

Similar definitions may be derived for other moral properties, such as *goodness* or *wrongness*, though I shall not do that here. Also note that in this thesis I shall restrict my attention to 'thin' evaluative terms and concepts, such as *right*, *bad*, and *ought*, setting aside 'thick' evaluative terms and concepts (which seem to additionally have a descriptive role), such as *cruel*, *courageous*, and *kind*.[10]

On my account, every possible ~*x* in (M_R) is, to some extent, *wrong* for *P* in those circumstances. However, *wrongness* exists on a continuum, so any particular ~*x* can be more or less wrong, with the degree of wrongness scaling according to the extent to which the action concerned would generate results that deviate (in terms of desirability for *P*, when fully rational and sufficiently informed) from the results of doing *x*.

Now, one may wonder what would be the consequence if we (like Gert himself) were to reject the common view that morality should never be overridden, holding instead that while moral behaviour is always rationally *permissible,* it is not always rationally *required*. One way to approach this question is to start with Gert's own conception of morality as a public system of imperatives that applies to all rational persons, governing behaviour that affects others (but which may rationally be overridden). Call this system (as a set of imperatives) Morality_G. Some of the imperatives in Morality_G may be those where the moral requirements in question are

[9] By a person's 'true strongest desire', I mean the strongest desire they would have when fully rational and sufficiently informed, as distinct from the strongest desire they might *presently* have.
[10] For more on *thick* terms and concepts, see: Simon Kirchin (ed.), *Thick Concepts* (Oxford: Oxford University Press, 2013).

also *rational* requirements. However, suppose that some, i.e. $\{i_1, i_2, \ldots, i_n\}$ are not. Now, replace each of these imperatives with the imperative that *would* be the rational requirement in those particular circumstances (assuming that there is always *some* rationally required behaviour), noting that these would then be the imperatives generated on my conception of morality, i.e. $\{i *_1, i *_2, \ldots, i *_n\}$. Call the new system Morality$_N$. Gert would presumably allow that there is a system Morality$_N$, even though he would not call it the 'moral' system. Then, to the extent that Morality$_G$ and Morality$_N$ come apart, Morality$_G$ would be irrelevant to the actual conduct of rational persons, with its imperatives being overridden for such persons by those in Morality$_N$. Accordingly, I submit that we should all then attend to Morality$_N$ instead of Morality$_G$, since it is the imperatives in Morality$_N$ that govern the behaviour of rational persons, with those in Morality$_G$ being overridden for them by those in Morality$_N$ whenever the two diverge. Therefore, to that extent, I think we should agree, with all of us attending primarily to what most people and I would call morality, with Gert's conception of morality then being consigned to practical irrelevance for rational persons.

Moreover, let me also consider how the previous positive argument for Goal Theory would run if we were to adopt a conception of morality that reflects Gert's view:

> **(Morality$_3$)**: an informal public system of imperatives that applies to all rational persons, governing behaviour that affects others, but which *may* be overridden.

On similar reasoning to before (but setting aside the details), we may then infer:

> **(S_1*)**: if there is a true moral system, then its system of imperatives may not supersede all other imperatives for rational persons.

Then:

> **(S_2*)**: if there is a true moral system, then its system of imperatives may not dictate what rational persons ought most to do.

And then:

> **(S_3*)**: if there is a true moral system, then its system of imperatives may not state what persons *will do*, when fully rational and sufficiently informed.

Thus, excluding the condition that morality should rationally never be overridden yields a moral system whose system of imperatives may include some that persons would only freely obey if they were (to some extent) irrational or ignorant. However, a moral system that includes imperatives that may correctly be disobeyed by all rationally informed persons does not sound like a kind of morality anyone should care about, but rather it sounds like a false morality. Anyone endorsing such a conception of morality would then presumably want to prevent people from becoming *too* informed and rational, lest they start acting ‘immorally’. However, this is a *prima facie* implausible result, further suggesting that Gert’s particular conception of overridable morality is defective. In any case, since it is also an uncommon conception, I shall set it aside.

Notice that because I am granting that morality should never be overridden, then these moral oughts will trump other oughts, e.g. prudential or legal ones (where these areas, like morality, are potential species of the genus of normativity). Moral oughts would then have the status of all-things-considered oughts (a *metanormative* claim). That said, observe that these moral oughts will plausibly align with prudential ones, insofar as acting so as to best serve the strongest desire of a fully rational and sufficiently informed version of oneself in such and such circumstances will also plausibly bring one the best consequences. Thus, on my account, prudence and morality do not appear to pull in different directions. I would further suggest that moral oughts so understood would routinely align with *legal* ones, insofar as the penalties incurred by acting illegally will often frustrate one's true strongest desire (though not universally, thereby allowing for morally permissible law breaking under exceptional circumstances).

Thus, I would argue that beginning with a definition of morality that most parties to the debate could assent to, and then proceeding through several steps that are at least plausible, we arrive at an account of the true moral system (if there is one), *T*. With reference to what I said at the beginning of the chapter, notice that, on *T*, morality is a system of *hypothetical* imperatives, not categorical ones. Now, let me turn to Goal Theory.

In line with a view notably expressed (though later repudiated[11]) by Philippa Foot, Richard Carrier's Goal Theory conceives of morality as a system of hypothetical imperatives.[12] In Carrier's case, this is a system of hypothetical

[11] On her later view, Foot thought that we ought to adopt the view that there are various kinds of considerations that generate reasons, with moral considerations being one set of considerations that a rational agent must take into account: Philippa Foot, *Natural Goodness* (Oxford: Oxford University Press, 2001).

[12] Philippa Foot, 'Morality as a System of Hypothetical Imperatives', *Philosophical Review,* 81 (1972), 305-15.

imperatives that supersedes all other imperatives, with ‘ought’ being the connector in these imperatives, and what we ought to do in any particular moral circumstances being dictated by what we would desire most to do if we were rational and sufficiently informed. This conception Carrier formulates as follows:

1. If you do *x*, *A* will happen; and if you do ~*x*, *B* will happen.
2. When rational and sufficiently informed, you will want *A* more than *B*.
3. If when rational and sufficiently informed you will want *A* more than *B* (and if *B*, then ~*A*; and if and only if *x*, then *A*), then you ought to do *x*.
4. Therefore, you ought to do *x*.[13]

Though Carrier does not explicitly do this, we may formulate from the preceding a conception of a moral fact on Goal Theory:

> **(F)**: It is a *moral fact* on Goal Theory that agent *A* ought to do some action *x* in circumstances *C* if, when rational and sufficiently informed, *A* would desire the results of doing *x* in *C* more than they would desire the results of doing ~*x* in *C*.

From this, we may then formulate a conception of moral rightness on Goal Theory:

> **(R)**: Some action *x* is morally *right* in circumstances *C* for agent *A* just in case it is a moral fact that *A* ought to do *x* in *C*.

[13] Carrier, 'Moral Facts Naturally Exist (and Science Could Find Them)', p. 335.

Now, observe something interesting here: the conceptions of moral 'oughts', moral facts, and moral rightness on Goal Theory correspond to (Ought), (M_F), and (M_R) inferred from Argument 1. Accordingly, I would submit that probably, if there is a *T*, then *T* = Goal Theory (i.e. if there is a true moral system, then probably that moral system is Goal Theory). Moreover, I shall argue in the next section that there actually *is* such a *T*, with Goal Theory's moral facts obtaining, and these facts being independent of actual people's beliefs or attitudes.

Notice two further things here. Firstly, on Goal Theory, I can answer the question of what makes true strongest desire satisfaction the all-important feature that determines whether actions are morally right or wrong: this is implied by the plausible definition of morality with which I began my positive argument, viz. (Morality$_2$). As such, I explain it by reference to something more basic (which is itself formulated by means of an appeal to the community's linguistic intentions), thereby not relying upon brute facts for which no further explanation is available.

Secondly, from Goal Theory's definition of the moral ought relation in (Ought2), we can see (contra 'Hume's Law', on which an 'ought' cannot be derived from an 'is'[14]), that a moral ought on Goal Theory (i.e. *P* morally *ought* to do *x* in moral circumstances *C*) *is* derived from a purely factual claim (i.e. doing *x* in *C* would best serve the strongest desire of a fully rational and sufficiently informed version of *P*). Thus, on my account, once we have the relevant facts about an agent's true strongest desire and about what will best serve this desire in such and such circumstances, then this settles how they ought to act. As such, I maintain that facts about what someone *ought* to do are not a separate and further issue from facts about what *is* the case (contra the non-naturalist).

[14] There is controversy about how to interpret Hume on this.

What are we to make of my positive argument? Well, I would concede that it amounts to only a *plausible*, rather than a knockdown, argument for Goal Theory. There is room for rational disagreement — something that I acknowledge by including the word 'probably' in the appropriate premises and conclusions of Argument 1. However, this might already be more than we have for some mainstream theories.[15] I find the argument to be rationally persuasive, but those who are antecedently disinclined to endorse Goal Theory as the true moral system (either because they favour an alternative moral theory, or because they think that there is no true moral system at all) may remain unconvinced. Certainly, if, upon investigation, Goal Theory seemed to be defeated by some valid objection, or failed to meet some applicable adequacy criterion, then this would count against the theory, perhaps to the extent of outweighing the confidence in Argument 1.

Conversely, if we discovered that Goal Theory plausibly resists these objections and satisfies all of the applicable adequacy criteria, then its claim to truth would be greatly strengthened. Even in the absence of Argument 1, if Goal Theory turned out upon investigation to be theoretically adequate and resistant to dominant objections, then I think it would still lay claim to being a theory worthy of serious consideration. Accordingly, I would suggest that undertaking such a critical evaluation ought to be advocated by all parties to the debate, since the results may give us good reason to either affirm or reject Goal Theory (with the former having potentially far-reaching consequences).

[15] For example, Keith Burgess-Jackson points out that utilitarianism is commonly thought to be *unprovable*, with Sidgwick, for example, admitting that his acceptance of utilitarianism was a mere 'intuition': Keith Burgess-Jackson, 'Taking Egoism Seriously', *Ethical Theory and Moral Practice*, 16 (2013), 529-42 (pp. 539-40).

2.2 The metaphysics

In light of the above-mentioned definitions of moral facts and moral rightness on Goal Theory, what positions does the theory adopt on the nature, constitution, and structure of moral reality? Before I find out, a short excursus on the natural-non-natural distinction is called for.

All parties to this debate — the reductive naturalists, the non-reductive naturalists, and the non-naturalists — agree that there are natural facts and properties and that there are moral facts and properties. (I am bracketing anti-realist views here since all of the relevant parties to the naturalism/non-naturalism debate reject these.) Naturalists think that moral facts and properties are metaphysically and epistemologically similar in all important respects to (other) natural facts and properties, whereas non-naturalists deny this.

To understand precisely what they disagree about, why this matters, and to position Goal Theory correctly in the debate, we need an appropriate characterisation of the *natural*. One widely accepted possibility is expressed thusly by Moore:

> By nature then I do mean and have meant that which is the subject matter of the natural sciences, and also of psychology.[16]

Russ Shafer-Landau offers a related characterisation:

> Naturalism … claims that all real properties are those that would figure ineliminably in perfected versions of the natural and social sciences.[17]

[16] Moore, *Principia Ethica*, p. 92.
[17] Shafer-Landau, *Moral Realism: A Defense*, p. 59.

Also, Derek Parfit says that:

> Some fact is natural, on one common definition, if facts of this kind are investigated or discussed by people working in any of the natural or social sciences.[18]

Moore's and Parfit's characterisations are silent as to whether they refer to the *current* objects and events of natural and social scientific and psychological investigation or those of *perfected* versions of these. Both options are problematic. Our natural and social scientific and psychological investigation is continually revising its ontology, and we do not (and possibly cannot) know which things outside the scope of our current natural and social scientific and psychological investigation would fall outside the scope of perfected versions.[19] In response to the problems inherent in such *disciplinary* characterisations, David Copp suggests an *epistemic* characterisation, whereby we should instead define the natural as being 'based in empirical observation and induction.'[20] However, that may be overly broad. Other characterisations have their own problems.[21]

However, when it comes to defending Goal Theory's particular conception of the relationship between the natural and the non-natural worlds, I shall bypass this debate by only referencing a subset of the natural that most people in this debate would agree is included in their definition. Specifically, whatever else nature includes, most relevant parties, irrespective of their metaphysical and

[18] Derek Parfit, *On What Matters Vol 2* (Oxford: Oxford University Press, 2011), p. 305.

[19] This is a version of Hempel's Dilemma. See, for example: C. Hempel, 'Comments on Goodman's Ways of Worldmaking', *Synthese,* 45 (1980), 193-99.

[20] D. Copp, 'Normativity and Reasons: Five Arguments from Parfit against Normative Naturalism', in *Ethical Naturalism: Current Debates,* ed. by Susana Nuccetelli and Gary Seay (Cambridge: Cambridge University Press, 2011), pp. 24-57 (p. 28).

[21] For example: R. Crisp, 'Naturalism and Non-Naturalism in Ethics', in *Identity, Truth and Value,* ed. by S. Lovibond and S.G. Williams (Malden, MA: Blackwell Publishers, 1996), pp. 113–29; D. Lewis, 'New Work for a Theory of Universals', *Australasian Journal of Philosophy,* 61 (1983), 343–77; M. Little, 'Moral Realism 2: Non-Naturalism', *Philosophical Books,* 35 (1994), 225–32. See also Carrier's definition: Richard Carrier, 'On Defining Naturalism as a Worldview', *Free Inquiry,* 30 (2010), 50-51.

epistemological commitments, would agree that it contains that which is the subject matter of our *current* natural and social science (where I take the latter to include psychology). With this understanding in place, I shall now return to the metaphysics.

Goal Theory is a *realist* view (and thence also a *cognitivist* one), positing moral facts for individual agents — as defined by (M_F) and (M_{F2}) from section 2.1 — with these facts being *objective* in the sense of being independent of actual people's beliefs or attitudes (since any such beliefs, or ways of thinking or feeling about something, have no bearing upon whether actions will or will not best serve agents' true strongest desires in such and such circumstances).[22] It is also *prima facie* a *naturalist* account, with moral judgements on Goal Theory rendered true or false by natural states of affairs (of human desire and cause and effect). But do moral facts really obtain on Goal Theory? And, if they do, are they *natural* facts (where negative answers to these questions would render Goal Theory anti-realist or non-naturalist, respectively)?

According to the definition (M_F), Goal Theory holds it to be a *moral fact* that *P* morally ought to do *x* in moral circumstances *C* if, when fully rational and sufficiently informed, *P* would desire the results of doing *x* in *C* more than they would desire the results of doing ~*x* in *C*. For any particular *P*, *x*, and *C*, the moral fact in question will therefore be composed of facts about what *P* would desire most, when fully rational and sufficiently informed (i.e. their true strongest desire), and facts about the outcomes of actions. Taking the latter component first, most parties would agree that facts about cause and effect obtain, including in cases where they

[22] On Goal Theory, moral facts are also *subjective* in the particular sense that they are defined in terms of what an (idealised) person would desire (most) to do. Thus, if one conceives of *subjectivism* as defining the content of moral judgements so that moral facts are about the subjective responses of moral appraisers and moral agents (where a relevant subjective response would include desire), then Goal Theory would also qualify as a subjectivist view. As such, on my conceptions, Goal Theory may be realist, yet simultaneously subjectivist.

relate to human behaviour and its outcomes, and that such facts are unproblematically natural, being part of the subject matter of our current natural and social science (and thus in accord with my understanding of the natural). In terms of the former component, I think that most parties would agree that facts about a person's desires obtain and that these facts — as psychological facts — are also part of the subject matter of our current natural and social science, and so should also be counted as unproblematically natural, being investigable and verifiable through such scientific disciplines as psychology, cognitive science, and so on. S*trongest* desires would also seem to be unproblematic. After all, we know that some desires are stronger than others, and can submit to others and derive from others. If there are stronger desires, then it is logically necessary that there must be one or more *strongest* desires (just as there must exist a tallest mountain), with these being in principle no less scientifically investigable than non-strongest desires. There is, however, a potential complication here, insofar as the strongest desire in question is that of a *fully rational and sufficiently informed* version of *P*. Does this make a difference? I would argue not. To see why, consider the following thought experiment.

Imagine ordinary person *P*, who has strongest desire *d*. In this case, I submit that facts about *d* obtain (e.g. that *d* = happiness) and that such facts are natural ones, being part of the subject matter of our current natural and social science. As such, I submit that there are natural facts about *d* that obtain. Now, suppose that some incremental change to *P* was to push them marginally in the direction of becoming a fully rational and sufficiently informed version of themselves, *P**. This change might, for example, be the acquisition of some relevant true belief, the destruction of some relevant false one, some slight improvement in their ability to make deductively valid or inductively forceful inferences, or the remedying of some cognitive bias from which they suffer. Let the improved version of P be called P_1, and their strongest desire at this point be called d_1 (where d and d_1 may be identical). Now, P_1 and d_1 can in practice exist, with the transition from P to P_1 being an everyday event (just providing P with some true belief would suffice in this regard). In that case, I submit that facts about d_1 obtain, and that these would be natural facts (with there being nothing about P_1 and d_1 — as compared to P and d — that would prevent these facts from obtaining and being part of the subject matter of our current natural and social science). Now, imagine that this process of incremental improvement is repeated n times. By increasing the value of n sufficiently, we can get as close as we want to the limit P^*, and at no definite point would facts about d_n no longer obtain, or obtain but no longer be part of the subject matter of our current natural and social science. As such, there is no definite point at which there would be no *natural* facts about d_n that obtain. Thus, I would argue that, in the limit, there are facts that obtain about the strongest desires of fully rational and sufficiently informed persons, with these facts being *natural* ones.

Accordingly, with the component facts from which Goal Theory's moral facts are composed plausibly obtaining and being natural, then, on the plausible assumption that facts composed solely of natural facts are themselves natural, it does indeed appear that there are moral facts on Goal Theory, and that these are natural facts. Therefore, I submit that my earlier claim that Goal Theory is a realist and naturalist view is vindicated.

In light of the above, we may identify Goal Theory as an *ethical naturalist* theory, where I shall understand ethical naturalism to be defined by the conjunction of two core theses: (1) that there are moral facts and properties; and (2) that these facts and properties are in an important sense *natural* facts and properties. Although ethical naturalism is distinct from and not entailed by metaphysical (or philosophical) naturalism, the latter nonetheless motivates the former.[23] Ethical naturalism fell out of favour for much of the last century, due in part to the influence of G.E. Moore's Open Question Argument (where he assumed the property of goodness is real, queried its nature, and ultimately argued that it is a *sui generis* non-natural property).[24] However, there is a renewed interest in the view today, at least partly because of doubts about the cogency of Moore's objections.

Moreover, ethical naturalism appeals because of its ability to offer a non-eliminativist account of morality that locates morality in the natural world. Further, in conceiving of moral facts and properties as natural facts and properties, it is generally agreed that the problem of supervenience dissolves. Thus, it is a plausible conjunction of two plausible views, viz. moral realism and naturalism. What is more, it holds on to the philosophically attractive thesis of representationalism about moral terms and sentences, whereby at least some moral terms denote legitimate natural properties, and some moral sentences represent how things are morally (implied by the ethical naturalist's view that at least some moral sentences have truth conditions of the sort countenanced by a robust moral realist theory). As such, Goal Theory is also a *representationalist* theory.

As we can see, Goal Theory does not identify moral facts with *irreducible* natural facts. Instead, it proposes that moral facts be *reductively* identified with complex natural facts composed of idealised (i.e. fully rational and sufficiently

[23] I shall understand *metaphysical naturalism* as the view that everything is composed of natural entities, of the kind studied by the (natural and social) sciences, and whose properties determine all of the properties of things, persons included. Abstracta like possibilia and mathematical objects, if they exist, would be constructed of such abstract entities as the sciences allow. I make no commitment here to metaphysical naturalism.

[24] Moore, *Principia Ethica*, pp. 66-68.

informed) human desire and cause and effect. As such, it is a natural *reductionist* account. However, is it best thought of as an *analytic* natural reductionist account or a *synthetic* one? This distinction can get worked out in a number of ways, but for my purpose I shall understand analytic natural reductionism as the position that some moral claims are semantically equivalent to certain non-moral claims (especially moral claims that state general relations between natural properties and moral properties, e.g. ‘pleasure is good’), with it being possible in principle to go through a process of conceptual analysis that would reveal these semantic equivalences. Synthetic natural reductionists deny such semantic equivalences, holding that all moral claims are synthetic ones, knowable by empirical methods.

In Goal Theory’s case, on definition (M_{R2}) from section 2.1, we have: *rightness* for person *P* in circumstances *C* just is the property that actions have when they best serve *P*’s true strongest desire in those circumstances. To put it another way, *rightness* for *P* in circumstances *C* can be reductively identified with being the best serving of *P*’s true strongest desire in those circumstances. So, has the reduction from the moral to the natural been secured by a process of conceptual analysis alone? If so, does this make my account an analytic one?

To take the second question first, I am not positing a *semantic equivalence* between rightness and some natural property, such that my definition of *rightness* in terms of best serving one’s true strongest desire captures the existing meaning of the term. Rather, I am proposing this as a possible replacement for our current vocabulary. Like the analytical naturalist, I hold that some moral properties (including *rightness*) are identical to natural properties. However, the analytical naturalists and I then part company, insofar as I (but not they) hold that the relation between the moral and the natural involves properties and facts only (making it a purely metaphysical relation), and so moral predicates and sentences are not content-equivalent to (and thereby not replaceable without significant loss by) purely descriptive predicates and sentences.[25] As such, even if the reduction from moral to natural properties were secured by conceptual analysis alone, I would still argue that mine would not be an analytic account in the ordinary sense.

Moreover, in terms of the first question, I would argue that the reduction has *not* in fact been secured by conceptual analysis alone. Although all of the steps leading to my definition of moral rightness are arguably conceptual ones, we have not reached the end of the analysis, because we still need to establish what *P*’s true strongest desire actually *is* — and that is a matter for *empirical* investigation. Before that final step, we merely have a *placeholder* for the natural property with which rightness is being reductively identified, not the natural property itself. Upon investigation, we might discover, for example, that *P*’s true strongest desire is for deep and abiding satisfaction. In that case, moral *rightness* for *P* in circumstances *C* just is the property that actions have when they best serve *P*’s desire for deep and abiding satisfaction in those circumstances. It is only at this point that we have fully secured the reduction from a normative property to a natural one; and because this final stage in the analysis is empirical, then I would once again argue that Goal Theory is not an *analytic* natural reductionist account in the ordinary sense.

At the same time, my account is also atypical of synthetic accounts, insofar as I do not attempt to justify the postulation of its moral facts and properties by arguing that they figure ineliminably in the best explanation of experience (as Cornell Realists do, for example), or claiming that the proposed identities are delivered by

[25] I take Jackson and Smith’s accounts to be representative of contemporary analytical naturalism: Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*; Michael Smith, 'Moral Realism', in *The Blackwell Guide to Ethical Theory,* ed. by H. LaFollette (Oxford: Blackwell, 2000), pp. 15-37.

means of an empirical hypothesis that is justified on the basis of enabling explanations of the relevant phenomena (as Railton does). In fact, I would suggest that we would suffer no explanatory loss in avoiding any reference to Goal Theory's moral facts and properties in explaining, e.g. why we would form the belief that someone acts wrongly if they pour petrol over a cat and sets fire to it. In such a case, I think natural facts about the act itself in conjunction with natural facts about the evolution and nature of our moral sense are sufficient to explain the formation of such a belief (I talk more about this in sections 2.5 and 3.4). To that extent, I would agree with Harman.[26] However, I avoid moral scepticism by further agreeing with Harman that the postulation of moral facts and properties would be justified if this was based upon a reduction of such facts and properties to some independently respectable (i.e., explanatorily efficacious) natural facts and properties (e.g. of psychology and cause and effect) — and then proposing just such a reduction. As such, because they have the causal powers that those natural facts and properties with which they are reductively identified do, I avoid any charges that my proposed moral facts and properties are causally inefficacious (and thus not genuine facts and properties[27]).

Accordingly, we might perhaps think of my account as a kind of analytic/synthetic hybrid that avoids some of the weaknesses of both views. In particular, by denying that some moral claims are *synonymous* with certain natural ones, then it is not vulnerable to standard versions of the OQA (as analytic accounts are).[28] And by not attempting to justify the postulation of its moral facts and

[26] Gilbert Harman, 'Ethics and Observation', in *The Nature of Morality: An Introduction to Ethics* (Oxford: Oxford University Press, 1977), pp. 3-10.

[27] On this point, see: J. Kim, 'The Myth of Nonreductive Materialism', *Proceedings and Addresses of the American Philosophical Association,* 63 (1989), 31-47.

[28] In section 6.4, I critically evaluate a contemporary version of the OQA that may be targeted at my account.

properties by arguing that they pull their weight in explanatory theories, then my account is not vulnerable to the charge (sometimes levelled at synthetic accounts) that there are no such moral facts and properties, on the basis that these facts and properties are causally inefficacious.

In light of the foregoing, I suggest that I have found a non-disjunctive way of picking out a naturalism-friendly (though unobvious) candidate for the referents of moral terms. Mine is a *revisionist* account, in the sense that the content of everyday judgements concerning moral rightness has to be revised in terms that appeal to the naturalistic property with which my account reductively identifies it. However, I would suggest that this revisionism is nonetheless *tolerable* — being analogous to the *vindicative* reduction of water to H_2O, rather than to the *eliminative* reduction of ‘polywater’ to ‘water that contains some impurities from improperly washed glassware’ (to use Railton’s examples), thereby reinforcing rather than impugning the sense that there really is such a moral property.[29]

My account bears some superficial resemblance to Railton’s reductionism, insofar as both might be thought of as reductive naturalist kinds of idealised subjectivism, with moral facts and properties being about the subjective responses (e.g. desires) of idealised moral agents. However, there are also substantive differences between the two accounts. I have already described how, on Railton’s account, but not mine, the identity for moral rightness is delivered by means of an empirical hypothesis that is justified on the basis of enabling explanations of the relevant phenomena. Beyond that, a significant distinction is that moral *rightness* on Railton’s account is constituted by ‘what is instrumentally rational from a social point of view’, yet Goal Theory denies such an impartial ‘social point of view’ where

[29] For more on vindicative and eliminative revisionism, see: Peter Railton, 'Naturalism and Prescriptivity', *Social Philosophy and Policy*, 7 (1989), 151-74 (pp. 159-61).

the interests of all potentially interested individuals are counted equally, adopting an *individual* point of view instead, on which the only morally relevant interests for an individual moral agent are the agent's own.[30]

2.3 The epistemology

Goal Theory locates the domain of morality within the familiar natural world, with its moral facts being composed of facts of idealised human desire and cause and effect. (For simplicity, I refer here only to moral *facts*, but the same applies to properties.) Accordingly, I would argue that Goal Theory's moral facts are in principle discoverable by the familiar methods of science, with no requirement to posit some special faculty (e.g. of intuition) or other means (e.g. reflection upon how things seem to us pre-theoretically) by which we may come to know *sui generis* moral facts (as the non-naturalist must do). Accordingly, moral knowledge is conceived of as just one more species of our knowledge of the natural world.

At this point, the critic may demur, denying that the (strongest) desires of *idealised* (i.e. fully rational and sufficiently informed) agents are in principle discoverable by the familiar methods of science. In response, let me construct a similar thought experiment to that from section 2.2. If the strongest desire, *d*, of ordinary person, *P*, can in principle be empirically specified and established by empirical investigation (as I would argue that it can), then is there anything in the incremental transition from person *P* to P_1 (i.e. making some slight improvement in rationality or knowledge) that would render d_1 not in principle empirically specifiable or establishable? I would argue not (and we could in principle take some

[30] Railton, 'Moral Realism', p. 200. I discuss this aspect of my account in more detail in chapter 5.

actual *P*, transform them into P_1, and test this claim). What about the transition from P_1 to P_2? Again, I would argue that there is nothing inherent in the transition that would render d_2 not in principle empirically specifiable or establishable. So, once again, imagine repeating this process of incremental improvement *n* times. By increasing the value of *n* sufficiently, we can get as close as we want to the limit *P** (i.e. the fully rational and sufficiently informed version of *P*), and at no definite point does d_n become not in principle empirically specifiable or establishable. Thus, I would argue that, in the limit, the strongest desires of fully rational and sufficiently informed agents are in principle empirically specifiable and establishable, and thus in principle discoverable by the familiar methods of science.

Of course, establishing Goal Theory's moral facts in practice may be a non-trivial undertaking. After all, the real world in general, and human beings in particular, are extraordinarily complex, meaning that (true) strongest desires and the consequences of particular actions may be exceedingly difficult to determine with any reliability. However, a proper scientific enquiry into Goal Theory's moral facts has yet to be implemented, and science has established much knowledge that was previously thought inaccessible (including in the fields of cosmology, particle physics, psychology, sociology, and cognitive science), despite facing considerable methodological difficulties. Just as science has verified facts in these and other fields, then I would expect that a suitable research program (e.g. in psychological and sociological science, and, eventually, neuroscience) would eventually yield results in Goal Theory's moral realm, with pessimism on this front being premature. Accordingly, I think that we require philosophy in order to specify correctly what moral facts *are*; but, having done this, we require only science in order actually to *discover* them.

In other words, Goal Theory’s moral facts are composed of empirical facts of ideal agents’ desires and the effects of actions. These facts seem, in principle, to be empirically specifiable and establishable, and science can discover any empirical facts that it develops methods capable of discovering. Moreover, there is much historical precedent for science developing methods capable of discovering empirical facts that were previously thought inaccessible. Thus, I think there is inductive reason to think that science could develop the required methods (at least to a sufficient degree), and thence discover moral facts on Goal Theory.[31]

Even if in practice we are never able to access perfect knowledge in this area, approximate knowledge of the necessary human psychology and cause and effect (and thus of putative moral facts on Goal Theory) is already accessible, and approximate knowledge is still valuable.[32] As Carrier says, ‘we needn’t know exactly what’s in an atom to make successful predictions from approximately what’s in an atom’.[33] Thus, even if we do not know the right thing to do, we can still know what the right thing is, *given what we know so far*; and that is optimal in the absence of perfect knowledge. Moreover, this potential limitation is not peculiar to Goal Theory, but instead befalls all ethical systems. For example, one might never know true Kantian categorical imperatives, because one cannot know enough about the world to correctly predict all potential contradictions in universalizing a rule; and one might never know the true utility maximising action because one cannot ever know the total causal outcome of every possible decision. However, a problem that befalls all moralities cannot be used as an objection to any particular one.

[31] For a formal argument making this case, see: Carrier, 'Moral Facts Naturally Exist (and Science Could Find Them)', pp. 363-64.

[32] On the value of approximate knowledge, and a defence of the principle that it is optimal in the absence of perfect knowledge, see: Walter Sinnott-Armstrong, *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*, 4 vols, Vol. 1 (Cambridge, MA: MIT Press, 2008), pp. 1-46; Kees Van Deemter, *Not Exactly: In Praise of Vagueness* (Oxford: Oxford University Press, 2010).

[33] Carrier, 'Moral Facts Naturally Exist (and Science Could Find Them)', p. 424.

2.4 The psychology

What, if anything, follows from making a sincere moral judgement on Goal Theory? In particular, do such judgements necessarily *motivate* (at least to some extent), or is any motivation only contingent?[34]

In order to answer these questions, consider first *A**, who is the fully rational and sufficiently informed version of moral agent *A*. Now, on Goal Theory, if *A** sincerely judges that she morally ought to do some act *x* in circumstances *C*, then she believes that doing *x* in *C* is the thing that will best serve her strongest desire in those circumstances. Accordingly, *A** has a pre-existing desire (i.e. her strongest desire) and a means-end belief about what action will best serve that desire in such and such circumstances (i.e. *x*). Thus, on the dominant Humean theory of motivation (HTM), where motivation requires the presence of a belief and an appropriately related and independently intelligible desire, *A** would then be *motivated* to abide by her moral judgement.[35] In fact, since the desire in question is *A*'*s *strongest* desire, then, on Goal Theory and the HTM, *A** will be *overridingly* motivated to act in accordance with her moral judgement.

Observe that, on Goal Theory and the HTM, this connection between sincere moral judgement and motivation for *A** to do *x* is a *necessary* one, since, by definition, the very circumstances of *A** making this judgement on Goal Theory entail the presence of a corresponding (strongest) desire and an appropriately related means-end belief. If the desire in question were absent, then there could be no

[34] When I say, motivate 'to some extent', I am suggesting a *disposition* to do such and such, where this disposition might be frustrated by circumstances, or overridden by other motivations.

[35] *A** may also be motivated on anti-Humean theories, where a belief is sufficient to motivate directly, or where the belief necessitates a desire, and the conjunction of the two motivates. However, I shall restrict my subsequent analysis to include the HTM only — both for simplicity and because I endorse the HTM (though I shall not argue for it here).

sincere moral judgement to do *x*. As such, whilst an appropriate desire is required for the moral judgement to motivate, on my account a sincere judgement necessarily entails the presence of such a desire. Therefore, when referring solely to fully rational and sufficiently informed agents, Goal Theory would be an *internalist* theory about motivational judgement (of the *weak* variety, on which there is a necessary connection between moral judgement and motivation; as opposed to the *strong* variety, on which moral judgment itself motivates, without the need for an accompanying desire).

Of course, real-world agents are not fully rational and sufficiently informed, so what result do we find on Goal Theory for these kinds of agents? Well, if *A* now sincerely judges that she ought to do some act *x* in moral circumstances *C*, then she is judging that doing *x* in *C* is the thing in those circumstances that would best serve the strongest desire of a fully rational and sufficiently informed version of herself (i.e. *A**). Now, if we call *A*'s desire set *D*, *A**'s desire set *D**, and *A**'s strongest desire *d**, the question of whether *A*'s sincere moral judgement to do *x* in *C* would necessarily motivate her to act accordingly seems to turn upon whether *d** is necessarily present in *D*. If *d* is* necessarily present in *D*, then *A* would necessarily be motivated to some extent do *x*, since *A* would then have a desire, *d**, and a means-end belief that doing *x* will serve that desire.[36] One way to approach this question is to consider the process by which the desire set might change from *D* to *D**.

Consider Smith's account of the belief-desire process, where there is an interaction between desires and beliefs through deliberation, and this interaction can generate new desires and destroy old ones. According to him, our fully rational self

[36] I say motivated 'to some extent' here, as *d** may be in *D*, but not be the *strongest* desire in *D* (unlike with *d** and *D**, where this is so, *ex hypothesi*). Of course, even if *d** is not present in *D*, *A* might have some *other* desire that would be satisfied to some extent by them doing *x* in *C*. However, I set that possibility aside.

would be a version of ourselves that has no false beliefs and all relevant true ones, does not suffer from weakness of will or suchlike, and that deliberates correctly (I take this to be analogous to my notion of a fully rational and sufficiently informed agent). By a process of correct deliberation on their desires and true beliefs (including evaluative beliefs about whether these desires are justifiable or not), this agent would have created new and destroyed old desires, until they are finally left with a set of desires that is entirely systematically justifiable, by which Smith means that they are maximally unified and coherent, and beyond reasoned criticism.[37] Due to the creation of new desires and the destruction of old ones, this desire set might be radically different from the set possessed by our non-ideal selves. Moreover, according to Smith, during this deliberative process, new desires can be generated where there was no antecedent desire as a premise.[38]

If we accept the above view, then *d** need not necessarily be present in *D*, as it may in principle have been generated as an entirely new desire during the process of deliberation. Thus, on this view, *A* need not necessarily be motivated to do *x* in *C* (since they may not have *d**, or any other desire that would be satisfied by doing *x*). As such, *A*'s moral judgement that they ought to do *x* would be only *contingently* connected to a motivation to do this.

Neil Sinhababu would concur with Smith that we might deliberate upon desires and beliefs, thereby generating new desires and destroying old ones. However, he argues, contra Smith (and Darwall), that desires can be changed as the

[37] Smith describes this process here: Smith, *The Moral Problem*, pp. 157-61. He further clarifies what he means by a set of desires being entirely systematically justifiable here: Michael Smith, 'In Defense of "the Moral Problem": A Reply to Brink, Copp, and Sayre-Mccord', *Ethics*, 108 (1997), 84-119 (p. 90).

[38] Darwall endorses a similar view: Stephen Darwall, *Impartial Reason* (Ithaca, NY: Cornell University Press, 1983), p. 39.

conclusion of reasoning *only* if a desire is among the premises of the reasoning.[39] On this view, the fact that *d** is a member of *D** implies that either *d**, or some antecedent desire, d_a, from which *d** was formed through suitable instrumental processes, must have been present in *D*. If not, then, on this view, *d** could not have been generated by a process of correct deliberation on desires and true beliefs. If *d** was in *D*, then *A* will necessarily be motivated (to some extent) to do *x* in *C*. Alternatively, if *d** was not present in *D*, but d_a was, then I would argue that, if doing *x* in *C* will serve *d**, then it seems plausible that doing *x* in *C* must also (to some extent) serve d_a (at least, on an account that takes the antecedent desire to be more general than the derived one). Thus, if there is a d_a, then it seems plausible that *A* must once again necessarily be motivated (to some extent) to do *x* in *C*.

Let me illustrate this with an example (originally from Darwall) that Sinhababu discusses (to rebut Darwall's claim that an agent forms a new desire through reasoning that does not have another desire as a premise).[40] Consider the case of Roberta, who has a desire — call it d_s — to promote a boycott [of goods from one company that has been particularly flagrant in its illegal attempts to destroy the union]. Now, on Sinhababu's account, Roberta must have had some antecedently existing desire, d_a, from which d_s was formed through some suitable instrumental processes. For example, we might say — as Sinhababu suggests — that d_a = relieve suffering. However, if we now suppose that d_s will be served by an action *x*, where *x* = donating a few hours a week to distributing leaflets at local stores, then, on Sinhababu's account, it seems plausible that doing *x* must of necessity also (to some extent) serve d_a.

[39] Neil Sinhababu, 'The Humean Theory of Motivation Reformulated and Defended', *Philosophical Review*, 118 (2009), 465–500.

[40] Sinhababu, 'The Humean Theory of Motivation Reformulated and Defended', pp. 482-89. The example is found here originally: Darwall, *Impartial Reason*, p. 39.

So, whether non-ideal agent *A* will necessarily be motivated (to some extent) to do *x* in *C* now appears to turn upon our preferred accounts of the belief-desire process. If desires can be changed as the conclusion of reasoning even if a desire is not among the premises of the reasoning (as Smith and Darwall claim), then the connection between moral judgement and motivation for *A* is a contingent one. Otherwise, it is plausibly a necessary one.

However, even if we were to argue that the connection between moral judgement and motivation for *A* is only a contingent one, if *A**'s strongest desire would be best served by doing *x* in *C*, then I would suggest that *A* very likely *will* be motivated (to some extent) to do the same. Firstly, and notwithstanding scepticism from Smith (who argues that *de dicto* desires to do what is right are fetishistic[41]), I think that many agents probably *do* have a general standing desire to do what they believe is morally right. Here I do not make the strong claim that moral agents are ultimately motivated only by a *de dicto* desire to do what is right, with all other relevant desires deriving from this (where, by a *de dicto* desire, I mean a desire that has a content that involves the concept of rightness — such that, if an agent is motivated by a desire *de dicto* to do what is right, then they desire to do this because it is right). Rather, I make the weaker claim that, amongst other (*de re* and *de dicto*) desires, many people will have a *de dicto* desire to do what is right, and they may derive some other (realiser) desires from this.[42] In that case, on Goal Theory, they

[41] In the context of assessing a possible externalist account of his so-called 'striking fact' (whereby 'a change in motivation follows reliably in the wake of a change in moral judgement'), Smith says: 'Good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality and the like, not just one thing [a general standing desire to do what is right] … Indeed, commonsense tells us that being so motivated is a fetish or moral vice, not the one and only moral virtue.' Smith, *The Moral Problem*, p. 75.

[42] I shall set aside further justification for this claim, but for arguments (contra Smith) that *de dicto* desires are not fetishistic, and that moral people are often motivated by both *de re* (e.g. for the welfare of loved ones) and *de dicto* desires, see, for example: Vanessa Carbonell, 'De Dicto Desires and Morality as Fetish', *Philosophical Studies: An International Journal for Philosophy in the Analytic*

may then derive from this general standing desire to do what they believe is morally right (i.e. *x* in *C*) a *pro tanto* (realiser) desire to do *x* in *C*. So, in conjunction with a means-end belief that doing *x* in *C* will serve this desire, they may then be motivated to some extent to do this.

Secondly, I think it is plausible that, upon suitable deliberation, most people would generate a *pro tanto* realiser desire to do what they know a fully rational and sufficiently informed version of themselves would do in those circumstances (based upon a general standing desire to act rationally and informedly). As such, in conjunction with the aforementioned means-end belief, they would once again be motivated to some extent to do *x* in *C*.

Thirdly, if (as I shall discuss in the next section) I am right to think that a suitable research program would determine that there is an (almost) universal true strongest desire amongst humans, and that this desire is for something in the region of a deep and abiding form of satisfaction (akin to Aristotle's eudaimonia), then it is hard to imagine that there could be many real-world agents (with normal psychology) for whom this putative true strongest desire would not be at least a *pro tanto* desire (and I think the same conclusion plausibly follows for other credible candidates for a universal true strongest desire too, e.g. pleasure, happiness, or truth).

I think that the conjunction of these three reasons makes it highly likely that real-world agents (with normal psychology) will have a means-end belief and an appropriately related desire to do *x* in *C*. Thus, I would argue that, on Goal Theory (and the HTM), even in the absence of any necessary connection, a real-world agent's moral judgement that they ought to do *x* in *C* will very likely generate a motivation to do this (albeit with the connection being a contingent one).

Tradition, 163 (2013), 459-77; Jonas Olson, 'Are Desires De Dicto Fetishistic?', *Inquiry,* 45 (2002), 89-96.

So, where does this leave us in terms of situating Goal Theory in the motivational judgement landscape? Well, we might describe Goal Theory as a kind of weak *internalist* theory, on the basis that there is a necessary connection between moral judgement and motivation for agents who are fully rational and sufficiently informed. This would then follow the example of Smith's moral rationalism[43], which he positions as an *internalist* theory.[44] Moreover, if we endorse Sinhababu's account of the belief-desire process, then even for real-world agents there is plausibly a necessary connection between moral judgement and motivation on Goal Theory — making it plausibly internalist on this basis too.

Alternatively, if we deny Sinhababu's account, then we might position Goal Theory as an *externalist* theory, on the basis that for all non-ideal agents (i.e. everyone in the real world), there will be only a contingent connection between moral judgement and motivation. However, even in this case, I would argue that there is good reason to suppose that almost everyone (of normal psychology) who sincerely makes a moral judgement will be motivated to some extent to comply with it; and the closer they approximate an ideal agent (as I understand this), the more reliably motivation will be connected to moral judgement, *ceteris paribus*. Thus, however we position it, Goal Theory can plausibly explain why it is that (almost) anyone who makes a sincere moral judgement would be motivated to some extent to comply with that judgement. As such, we might say that Goal Theory's definition of moral rightness captures the *practical nature* of morality, with moral judgements

[43] On which an agent will always be motivated to do what they believe is right, unless they are practically irrational, because for an agent to judge that doing *x* in *C* is right *is* for them to judge that if they were fully rational they would desire to do *x* in *C*.

[44] Adina Roskies, however, thinks the implied kind of internalism (which ties internalism to practical rationality) would be trivial and uninteresting (unlike an internalism on which moral beliefs or judgements are intrinsically motivating). A. Roskies, 'Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy."', *Philosophical Psychology*, 16 (2003), 51-66 (p. 52).

motivating those who make them. This result will be important in chapter 6, when I evaluate Goal Theory's theoretical adequacy.

As a cognitivist, Humean, and (on some understandings) internalist account, how would Goal Theory cope with Michael Smith's so-called 'moral problem'? As Smith points out, there are three intuitively plausible features of moral judgement that are in tension, with the acceptance of any two counting against the third. According to Smith, this tension explains the extent and nature of disagreement that we find in contemporary metaethics. Smith formulates this in terms of the following three propositions:

> (1) Moral judgments of the form 'It is right to Φ' express a subject's beliefs about an objective matter of fact.
>
> (2) If someone judges it right that she Φs, then, other things being equal, she is motivated to Φ.
>
> (3) An agent is motivated to act in a certain way just in case she has an appropriate desire and means-end belief, where belief and desire are, in Hume's terms, distinct existences.[45]

The moral problem is then that these three propositions are each intuitively plausible, but appear to be collectively incompatible (albeit not strictly inconsistent). The first of these propositions (a statement of cognitivism) Smith calls the 'objectivity thesis'; the second (a statement of internalism) he calls the 'practicality requirement'; and the third the 'Humean psychology'. To illustrate the tension between them, consider, for example, that if we affirm the first and third propositions, then it becomes hard to see how beliefs, which are deemed conceptually distinct from the conative states

[45] Smith, *The Moral Problem*, p. 12.

required for motivation, could themselves guarantee motivation (other things being equal), as the practicality requirement dictates. Alternatively, if we affirm the first and second of the propositions, then it seems that some beliefs are motivational, even in the absence of the independent psychological desire state posited by Humean psychology**.**

Smith says that any successful metaethical theory must either find a way to justify and successfully reconcile these three propositions, or else it must deny one of them and bite whatever bullet is thereby entailed. Accordingly, non-cognitivists would deny the objectivity thesis, holding instead that moral judgements express attitudes, emotions, or suchlike; externalists would deny the practicality thesis, claiming that whilst beliefs might be associated with motivations, this is not because there is some internal and necessary connection between the two; and anti-Humeans deny Humean psychology, asserting, for example, that at least some beliefs can be motivating in the absence of conceptually independent desires.

So, how would Goal Theory respond to this problem? Well, remember, on Goal Theory, if agent *A* judges that it is right for her to do *x* in *C*, then she is expressing a belief that her doing *x* in *C* is the thing that would best serve the strongest desire she would have if she were fully rational and sufficiently informed. Accordingly, Goal Theory would endorse Smith's (cognitivist) proposition (1). Secondly, I have argued that, depending upon our preferred account of the belief-desire process, if *A* makes the aforementioned sincere moral judgement, then she will either necessarily, or contingently but very probably, possess the desire that is her fully rational and sufficiently informed self's strongest desire (i.e. the one that will be best satisfied by her doing *x* in *C*). But this means that *A* would then at least very probably be motivated to some extent to do *x*, since she would have a desire and a means-end belief that doing *x* will serve that desire. Accordingly, Goal Theory would also yield a qualified endorsement of Smith's (internalist) proposition (2). This qualified endorsement may be sufficient for the purpose at hand.

So, in ostensibly affirming Smith's first two propositions, does it then follow that I must deny Smith's Humean proposition — something that I have previously endorsed? I would argue not. In Goal Theory's case, it is not that the belief expressed in proposition (1) is motivational *in and of itself*, in the absence of the independent psychological desire state posited by Humean psychology. Instead, as I have explained, on Goal Theory I think the nature of the belief expressed in proposition (1) (i.e. about doing *x* best serving *A*'s true strongest desire) is such that agent *A* will at least very probably always possess an appropriately related and independently intelligible desire — thereby at least very probably generating the motivation referred to in proposition (2). Thus, I would argue that Goal Theory deflates Smith's moral problem, plausibly reconciling his triad of propositions — entailing, explaining, or being compatible with each of them.

2.5 Egoism

Goal Theory claims that what one ought to do in such and such moral circumstances is effectively a function of what behaviours produce the best consequences. As such, it is a variety of *consequentialist* theory, denying the deontologist's claim that some actions are *inherently* right or wrong, regardless of the consequences of those actions. However, Goal Theory differs from utilitarianism with regard to which consequences are morally significant. In the latter case, it is the consequences for everyone in aggregate; in the former case, it is the consequences for the moral agent themselves. Accordingly, Goal Theory is a variety of *ethical egoist* theory (and Carrier explicitly positions it as such[46]).

By their nature, Goal Theory's moral facts are facts for *individual* moral agents — defined in terms of individual agents' true strongest desires. So, in principle, distinct moral facts might obtain for different moral agents in the same circumstances (with, for example, *x* being what *A* ought to do in circumstances *C*, but some ~*x* being what *B* ought to do in *C*). As such, Goal Theory is in principle a realist version of moral relativism. Whether in practice these moral facts differ from moral agent to moral agent is an empirical matter, but I submit that we will much more likely discover that moral facts on Goal Theory are (almost) universal.

Specifically, for reasons to do with our shared fundamental (as opposed to incidental) *biology* (e.g. everyone being members of the same species, with the same origins, and continued interbreeding), *conscious experience* (e.g. everyone constructing a conscious self-awareness when healthy and awake, everyone having mirror neurons, and everyone relying upon an innate theory of mind to understand others [or else learning and applying such a theory, as most autistics can do, for example]), and *environment* (both physical and social, with everyone needing to eat, sleep, move, breathe, think, cooperate with a social group, and avoid the same physical and emotional harms, for example), I think that upon suitable empirical investigation we will probably find that (almost) all agents would have the *same* true strongest desire, with the *same* actions best serving this desire in the same circumstances. Hence, if *x* is the thing on Goal Theory that agent *A* morally ought to do in circumstances *C*, then I submit that *x* will also very probably be what *B* morally ought to do in *C*.

[46] Carrier, *Sense and Goodness without God: A Defense of Metaphysical Naturalism*, p. 316.

I am prepared to go further out on a limb here and suggest a plausible candidate for this (almost) universal true strongest desire, viz. a form of deep and abiding satisfaction (something more than mere happiness or pleasure). Even before any empirical investigation, we already have some support for this claim, based upon an argument of Aristotle's. Adapting his argument to my needs, we have the following. All desires have a reason. We do not just desire things for no reason. Most things we desire, we desire because we desire something else that is achieved by it. Pick any desire, and ask why we want that, and we will realise that there is a reason to want that thing, a reason to have that desire. We desire it for some particular end, and not just for itself. Otherwise, we would not want it. Of course, this cannot go on forever. We do not have infinite desires. Therefore, there must be something (possibly many things) that we desire for no reason. However, Aristotle argued that there was one ultimate reason that we desire anything at all, and it is that singular state of satisfaction, of *eudaimonia*.[47] That, he said, is the only thing we do not desire for some other end, the only thing that we desire solely for itself. When we ask, 'Why do I want to be ultimately satisfied?', the question is inherently absurd. It would be like asking 'What is north of the North Pole?' There are of course different degrees of satisfaction (it can be measured qualitatively: some states of satisfaction are more desirable than others, and quantitatively: how often and for how long), and the greatest state of satisfaction, that than which no state is more satisfying, is perfect happiness. All lesser states of satisfaction are degrees of happiness, and we always aim at getting higher up that ladder, or in greater quantities. Our greatest goal in everything we do is simply this: the highest state of satisfaction we can obtain, for as long or as often as possible. This is not our present desire, but our true desire. That is, if we rationally and informedly reflected upon why we want anything, why we have

[47] Here I understand the satisfaction-state to be a particular psychological experience.

any desire we do, why we prefer anything to anything else, we will always come around to the same conclusion: because it satisfies us to do so. Pick any desire we have that motivates us, in fact, pick any strongest desire we have (a true desire, not a merely present desire) and ask ‘Why should I want that, rather than something else instead?’ The reason will always be some reference to the state of satisfaction we will obtain by realising that desire (and sometimes even by merely having the desire). Thus, our ultimate goal is that ‘satisfaction-state’. All desires are pursued for that end. Therefore, on the assumption that an intrinsic desire will be stronger than any desire to realise that intrinsic desire, then the intrinsic desire for eudaimonia is then our true strongest desire.

While I find this suggestion to be plausible, the truth of Goal Theory is not dependent upon confirming empirically that the proposed (almost) universal true strongest desire is a form of deep and abiding satisfaction — so Goal Theory is not made a hostage to fortune. Other than perhaps some diminishment in its degree of conservatism, Goal Theory would survive if the (almost) universal true strongest desire turned out to be something else (e.g. pleasure, happiness, or truth), or if there turned out to be *no* universal true strongest desire, with true strongest desires varying from agent to agent.

Given that I understand the satisfaction-state in question to be a psychological experience, one might wonder if my account is then vulnerable to Robert Nozick’s ‘experience machine’ challenge (which is commonly targeted at ethical hedonism).[48] To find out, imagine that we could obtain maximal deep and abiding satisfaction (as a psychological state) by plugging ourselves into the so-called experience machine for the rest of our lives, with the machine giving us experiences of whatever sort

[48] Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), pp. 43-45. Also: Robert Nozick, *The Examined Life* (New York: Simon & Schuster, 1989), pp. 104-08.

would produce this maximal satisfaction for us, without us being aware that we were on the machine, with there being no concern of it breaking down, no negative consequences for family and friends, and so on. Other people could also plug-in, so there would be no need to stay unplugged to serve them; and those plugged may have access to a common ‘virtual world’ shared by other machine-users (in which ‘ordinary’ communication is possible). In such hypothetical circumstances, would and should we plug in? Nozick thinks not, saying that ‘we want to *do* certain things, and not just have the experience of doing them’, that ‘we want to *be* a certain way, to be a certain sort of person’ (and not just an indeterminate blob floating in a tank), and that we do not want to be limited ‘to a man-made reality, to a world no deeper or more important than that which people can construct’.[49]

What does this challenge show? If Nozick is right that we would not plug into the experience machine, then we might conclude that there are things that we want besides deep and abiding satisfaction (in particular, an actual connection to reality). However, even if it is true, this conclusion is consistent with my account, since I do not claim that the *only* thing we want is deep and abiding satisfaction. One might also draw a stronger conclusion, viz. that we want this connection to reality *more* than we want deep and abiding satisfaction, meaning that deep and abiding satisfaction is then not our strongest desire. However, this conclusion is also consistent with my account, since I do not claim that deep and abiding satisfaction is our strongest *present* desire. Rather, I claim that it would be our strongest desire *if* we were fully rational and sufficiently informed (which none of us is). Therefore, even if it were true that we presently desire this connection to reality more than we desire deep and abiding satisfaction, then this would still not undermine my account.

[49] Nozick, *Anarchy, State, and Utopia*, pp. 43-44.

Now, Nozick might claim that we would not plug into the experience machine, *even if* we were fully rational and sufficiently informed, since an overriding desire to live an authentic life would persist for us even in this enlightened state. However, such an unsupported claim would be highly speculative, and would simply beg the question against my account. Moreover, I think we already have good reason to be sceptical of this claim. Let me explain.

When Nozick declared that we would not plug into the experience machine, he did not back up his claim with any empirical evidence. However, others have subsequently tested his claim empirically, broadly verifying it. For example, in one study, Dan Weijers found that only 16% of participants would permanently connect to the experience machine.[50] However, some were concerned that people's responses are subject to a number of biasing factors, and that this is generating a distorted result.[51] In response to such worries, Weijers reformulated Nozick's vignette in a way that tries to neutralise a number of these factors. In particular, he tried to neutralise biasing factors due to an overactive imagination ('the machine seems scary or unnatural'), imaginative resistance ('the machine might break down or not produce great experiences in the future,' 'unpredictable or surprising experiences are better than pre-programmed ones,' 'bad experiences are required to appreciate good experiences or to develop properly', and 'I can't because I have responsibilities to others'), loss aversion, and status quo bias. He found on his revised scenario that acceptance rates went up from 16% to 55%. Thus, when we correct for certain biases and other irrelevant factors, we find a majority in favour of plugging-in, notwithstanding any desire to live an authentic life.

[50] D. Weijers, 'Nozick's Experience Machine Is Dead, Long Live the Experience Machine!', *Philosophical Psychology*, 27 (2014), 513–35 (p. 520).
[51] E.g. L. W. Sumner, 'Welfare, Happiness, and Pleasure', *Utilitas*, 4 (1992), 199–223 (p. 216).

Filipe De Brigard also conducted studies on Nozick's thought experiment, where he attempted to neutralise status quo bias by formulating *inverted* experience machine scenarios, in which participants were told to imagine that their lives to date had been in the experience machine, and then asked whether they wanted to *disconnect*. He found that only 41% wanted to do so.[52]

In light of the above results, I think we have good reason to be sceptical of the claim that we would not plug into the experience machine if we were fully rational and sufficiently informed (where it is this claim that needs to be plausibly true in order to undermine my suggestion that our true strongest is for a deep and abiding satisfaction). This is especially so when we consider that not all biases and irrelevant factors could be neutralised in Weijers and De Brigard's studies, that people's responses might be subject to other failures of reason or absences of relevant information, and that fully rational and sufficiently informed people would presumably be subject to none of these issues.

Moreover, *ex hypothesi*, we do not know when we are plugged-in to the experience machine, so, while connected, we must have experiences of really doing things, really being a certain way, and not being limited to a man-made reality, that are indistinguishable to us from the real thing.[53] Moreover, to the extent that this is conducive to maximising our deep and abiding satisfaction, we could interact with other users in a common 'virtual world' (in addition, of course, to the virtual people fabricated by the machine). Thus, from our first-person perspective, where there is only our conscious *experience* of our own mental states (e.g. thought, memory,

[52] F. De Brigard, 'If You Like It, Does It Matter If It's Real?', *Philosophical Psychology,* 23 (2010), 43–57.

[53] Here I assume (in the spirit of the thought experiment) that the experience machine generates such a rich and authentic experience that it is indistinguishable from the real world to a person whose real-world memories and critical faculties are intact, setting aside the possibility that it generates a poor and inauthentic simulation of the real world but limits our real-world memories and critical faculties so that we cannot tell.

emotion, and desire) and the world around us (via the senses), then nothing goes missing when we are plugged-in as opposed to experiencing the real world. At the same time, when we are plugged-in, our deep and abiding satisfaction is maximised. Thus, under the conditions of the thought experiment, being plugged-in seems to improve upon not being plugged-in for us, accruing all of the benefits (including of perceived authenticity) and more (i.e. maximal satisfaction), without the drawbacks associated with living in the real world.

Despite this, we may nevertheless feel some queasiness about plugging-in. However, I would suggest that this is then due to some cognitive bias or other irrationality, to which, by definition, our fully rational and sufficiently informed selves would not fall prey.

As to whether we *should* plug into the experience machine: if, as I suggest, a fully rational and sufficiently informed person would probably do so (under the idealised conditions of the thought experiment), then I would answer in the affirmative, as to do otherwise would be to do what we would only (freely) do if we were irrational or insufficiently informed. (Of course, the idealised conditions of the thought experiment may never obtain in the *real* world — where the machine *might* break down, the operators of the machine *might* really be sadistic thrill-seekers, there *might* be negative consequences for family and friends, and so on — so I am not claiming that we ought to connect to the experience machine in anything other than the hypothesised conditions.)

Accordingly, I would argue that Nozick's thought experiment probably fails to undermine my claim that the (almost) universal true strongest desire is for a form of deep and abiding satisfaction.

Setting Nozick aside, if we grant that there is an (almost) universal true strongest desire and that this desire is for something in the region of a deep and abiding kind of satisfaction, then what does this suggest about the kinds of moral propositions implied on Goal Theory? Are they likely to include intuitively false ones (e.g. ones that command selfish, dishonest, and malevolent actions), as some critics of ethical egoism would insist? I would argue not. I will return to this in section 5.4, but for now I would suggest that, from empirical enquiry, in addition to the theoretical and practical application of game theory, we find that certain ways of acting (e.g. cooperatively, altruistically, and honestly) are generally in our best long-term enlightened self-interests (including in the promotion of our deep and abiding satisfaction) in social groups like ours where our interests are affected by what other people do, as well as what we ourselves do, and where everyone pursuing their individual short-term interests will make them all worse off.[54] This is due in one part to the benefits that accrue to us from other's direct and indirect reciprocity when we act altruistically, cooperatively, and compassionately towards them (with these benefits reliably enhancing our happiness and satisfaction); and in another part to the evolved psychological payoffs that we obtain from acting in this way (e.g. greater happiness and sense of belonging, and a reduction in feelings of isolation, stress, and negativity). Hence, I suggest any short-term loss incurred by cooperating with others (e.g. keeping an inconvenient promise) will generally be outweighed by a long-term gain (e.g. having one's future promises trusted). As such, I think that these types of behaviours plausibly serve the sort of true strongest desire adduced earlier.

[54] In terms of the beneficial game-theoretic effects of acting altruistically and cooperatively towards others (and vice versa), see, for example: Ken Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005). Also: Gary L. Drescher, *Good and Real: Demystifying Paradoxes from Physics to Ethics* (Cambridge, MA: The MIT Press, 2006), pp. 273-320. On the best strategies for infinitely repeated Prisoners' Dilemma scenarios in particular (viz. *cooperative* ones), see: Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984). And in terms of the positive emotional rewards of cooperation and altruism (and vice versa), see: Stephen G. Post, 'Altruism, Happiness, and Health: It's Good to Be Good', *International Journal of Behavioral Medicine,* 12 (2005), 66–77.

Conversely, when we act towards others in, say, selfish, dishonest, and malevolent ways, then we not only forego the aforementioned benefits, but we potentially incur harms, including sanctions, ostracism, retribution, and punishment. Moreover, we may also suffer adverse psychological effects, including stress, fear of capture, and feelings of self-loathing, isolation, guilt, and shame. None of this is conducive to serving over the long term anything like the candidate true strongest desire I propose (quite the opposite in fact). Therefore, I would submit that the kinds of actions that will be right or wrong for (almost) all agents on Goal Theory (which we might call *universal moral rules*) will generally align with what might be called *stock moral truisms* (understood as moral propositions that are widely accepted and intuitively true), e.g. that it is morally wrong to kill another person simply because it gives one pleasure.[55]

Psychopaths are sometimes posited as possible exceptions to this, being people whose strongest desires are ostensibly served by acting in sometimes highly immoral ways, and not served by acting morally. However, psychopaths act so self-defeatingly, and are so routinely dissatisfied with themselves and the world, that it can hardly be claimed that their behaviour best serves their long-term enlightened self-interest. They are just too irrational or ill-informed (or indeed, insane) to understand that fact.[56]

[55] Cuneo identifies some others, including that it is wrong to lie to one’s spouse simply to save face, and that it is wrong to break a promise simply because one feels like it: Terence Cuneo, 'Moral Naturalism and Categorical Reasons', in *Ethical Naturalism: Current Debates,* ed. by Susana Nuccetelli and Gary Seay (Cambridge: Cambridge University Press, 2011), pp. 110-30 (p. 118).

[56] For more on this, see: Carrier, 'Moral Facts Naturally Exist (and Science Could Find Them)', pp. 352-53. See also: Carrier, *Sense and Goodness without God: A Defense of Metaphysical Naturalism*, pp. 342-44. Additionally: Sinnott-Armstrong, *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*, p. 390. And: Walter Sinnott-Armstrong, *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, 4 vols, Vol. 3 (Cambridge, MA: MIT Press, 2007), pp. 119-296, 363-66.

2.6 Theory of normative reasons

The Humean Theory of Reasons (HTR) has been endorsed or taken for granted by many influential ethical naturalists, and may be seen as a natural fit with Goal Theory (given the fundamentally instrumentalist nature of moral imperatives on the theory).[57] For example, Peter Railton says of the HTR that it is:

> …the clearest notion what it is for an agent to have reasons to act. Moreover, it captures a central normative feature of reason-giving, since we can readily see the commending force for an agent of the claim that a given act would advance his ends.[58]

Furthermore, Richard Boyd says that:

> Ordinary factual judgments often provide us with reasons for action; they serve as constraints on rational choice. But they do so only because of our antecedent interests or desires. If moral judgments are merely factual judgments, as moral realism requires, then the relation of moral judgments to motivation and rationality must be the same.[59]

Other ethical naturalists who embrace the HTR include David Brink and Mark Schroeder.[60] Those who endorse it may proceed from the claim that reasons must be

[57] For more on the HTR, see, for example: Peter Railton, 'Humean Theory of Practical Rationality', in *The Oxford Handbook of Ethical Theory*, ed. by David Copp (Oxford: Oxford University Press, 2006), pp. 265-81.
[58] Railton, 'Moral Realism', p. 166.
[59] Boyd, 'How to Be a Moral Realist', p. 186.
[60] Brink, *Moral Realism and the Foundation of Ethics*; Mark Schroeder, *Slaves of the Passions* (Oxford: Oxford University Press, 2007).

able to motivate, add that motivation requires desire, and then conclude that having a reason requires having a desire. It may be stated as follows:

> **HTR**: an agent has *pro tanto* normative reason for an action just in case that action would serve some of the agent's desires.

However, Goal Theory is not committed to, or even compatible with, the HTR in the form described. Before I explain why, I must introduce a couple of ideas. Firstly, we have the connection between 'oughts' and normative reasons contained in the following statements, as expressed by Derek Parfit:

> If our reasons to act in some way are stronger than our reasons to act in any of the other possible ways, then these reasons are decisive, and acting in this way is what we have most reason to do…
>
> When we have decisive reasons, or most reason, to act in some way, this act is what we should or ought to do in what we can call the decisive-reason-implying sense.
>
> …[We may] have sufficient reasons, or enough reason, to act in any of two or more ways. Our reasons to do something are sufficient when these reasons are not weaker than, or outweighed by, our reasons to act in any of the other possible ways…[61]

Henceforth, I shall refer to these as 'Parfit's platitudes', and shall affirm them.[62] In line with these platitudes, I shall claim that agent *A ought* to do *x* (in circumstances *C*) just in case there is decisive (or sufficient) normative reason for *A* to do *x* in *C*.

[61] Derek Parfit, *On What Matters Vol 1* (Oxford: Oxford University Press, 2011), pp. 32-33.

[62] In a paper discussing Joyce's arguments for categorical reasons, Andrés Carlos Luco identifies these statements as having the status of platitudes in common normative discourse: Andrés Carlos Luco, 'Non-Negotiable: Why Moral Naturalism Cannot Do Away with Categorical Reasons', *Philosophical Studies*, 173 (2016), 2511–28 (p. 2524).

Secondly, we have the thesis of *Proportionalism*, according to which, when a reason is explained by a desire, its weight varies in proportion to the strength of that desire, and to how well the action promotes that desire.[63] I shall also endorse this thesis.

Now, on the conjunction of the HTR, Parfit's platitudes, and Proportionalism, it is possible that what an agent ought to do will come apart from what they ought to do on Goal Theory — since what will best serve an agent's strongest desire need not be what will best serve the strongest desire they would have when fully rational and sufficiently informed. As such, upon pain of incoherence, I must give up one or other of these (remembering from section 2.1 that Goal Theory's moral ought is an all-things-considered one, and thus should align with what we have decisive reason to do). In light of this, I would like to keep Parfit's platitudes and Proportionalism, but propose the following variant of the HTR, which I think improves upon the one above in several important ways, as I shall show in the next chapter:

> **HTR***: an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent.

[63] Mark Schroeder says that Proportionalism is almost universally thought to go along with the HTR, although he actually denies it, replacing it with his own weighting scheme for reasons: Schroeder, *Slaves of the Passions*, pp. 97-103. However, I think that Schroeder's weighting scheme is problematic. Amongst other things, as pointed out by Tristram McPherson, it may generate an *explosion* of agent-neutral reasons: Tristram McPherson, 'Review: Mark Schroeder's Hypotheticalism: Agent-Neutrality, Moral Epistemology, and Methodology', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 157 (2012), 445-53. Schroeder acknowledges that this is a potentially catastrophic problem for his account, and is unsure whether it can be resisted: Schroeder, *Slaves of the Passions*, p. 472.

We might describe the HTR* as a *counterfactual* version of reasons *internalism*, on which, if an agent has a reason to do *x*, then it follows by necessity that the agent would desire to do *x* if they were fully rational and sufficiently informed.

Notice that we do not get a conflict now. In fact, we can see that the HTR* follows directly from a conjunction of Goal Theory, Proportionalism, and Parfit's platitudes. Specifically, if *A* morally ought to do *x* just in case doing *x* will best serve the strongest desire of a fully rational and sufficiently informed version of *A* (as Goal Theory states), and if *A* ought to do *x* just in case there is decisive (or sufficient) normative reason for *A* to do *x* (as the platitudes say), then it follows that there is decisive (or sufficient) normative reason for *A* to do *x* just in case doing *x* will best serve the strongest desire of a fully rational and sufficiently informed version of *A*. (That is, if the following two propositions are both true: (1) if *x* ought to do Φ just in case *y*; and (2) if *x* ought to do Φ just in case *z*; then it follows that *y* = *z*.) However, if we then reduce the strength of the desire in question, and how well the action promotes that desire, then, on Proportionalism, *A* will have a *pro tanto* reason to do *x* just in case doing *x* will serve some of the desires of a fully rational and sufficiently informed version of *A* (as the HTR* says). Hence, if we take Goal Theory as our theory of morality, and endorse Parfit's platitudes and Proportionalism, then it follows that the HTR* will necessarily be our theory of normative reasons. Given the plausibility of Proportionalism and Parfit's platitudes, I shall therefore regard the HTR* as being Goal Theory's implied theory of normative reasons.

It would be possible to endorse the HTR* without endorsing Goal Theory, but observe that if one endorses the HTR*, Proportionalism, and Parfit's platitudes, then Goal Theory (or something like it) is thereby entailed. Specifically, if one maximises the strength of the desire referenced by the HTR*, as well as how well the action promotes that desire, then, on Proportionalism, one generates a decisive reason for an agent to act accordingly. And, according to Parfit's platitudes, this is then what they *ought* to do. Thus, we derive the proposition that what an agent ought to do is what will *best* serve the *strongest* desire of a fully rational and sufficiently informed version of themselves — just as Goal Theory specifies.

In light of the foregoing, if it were to turn out that there are *categorical* normative reasons — obtaining for an agent regardless of any true strongest desire they may possess — then the HTR* would be defeated. If the HTR* is defeated, then, granting Parfit's platitudes and Proportionalism, Goal Theory would thereby be undermined. This observation will be relevant in the next chapter.

Moreover, notice that because on Goal Theory the moral action in such and such circumstances for an agent is the one that best satisfies the agent's true strongest desire in those circumstances, then, on the HTR* (where an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent), agents will have excellent reasons for compliance with moral requirements (*decisive* ones, in fact, on Proportionalism and Parfit's platitudes).

Notice also that with my concept of a normative reason in place, I am in a position to provide an analysis of my concept of rightness in terms of my concept of a normative reason (something that Smith also attempts to do on his account[64]). Specifically, from section 2.1, my conception of rightness is as follows:

> **(M_{R2})**: moral *rightness* for *P* in moral circumstances *C* just is the property that an action has when it best serves *P*'s true strongest desire in those circumstances.

[64] Smith, *The Moral Problem*, pp. 182-203.

And my conception of a normative reason is the following:

> **HTR***: an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent.

Bearing in mind the notion of Proportionalism, the following conception of rightness is then implied on my account:

> **(M_{R3})**: moral *rightness* for *P* in moral circumstances *C* just is the property that an action has when *P* has *decisive normative reason* to do that action in those circumstances.

Moreover, incorporating Parfit's platitudes then allows me to rewrite this as follows:

> **(M_{R4})**: moral *rightness* for *P* in moral circumstances *C* just is the property that an action has when *P ought* to do that action in those circumstances.

I think this neatly illustrates the internal consistency in my account. Even though I am now incorporating my concept of a normative reason, as well as Proportionalism and Parfit's platitudes, I have derived the same relationship between moral *rightness* and what one *ought* to do that I might have derived by working directly from my definitions (Ought_2) and (M_{R2}) in section 2.1.

2.7 Questions and challenges

Having described Goal Theory's metaphysics, epistemology, psychology, egoism, and theory of normative reasons, I shall now turn my attention to some questions and challenges. Of course, given the obvious limitations in scope, many questions and challenges must be set aside, but I have selected those that I deem to be the most common or important. Some of these I shall answer quickly, but three will demand a more extended treatment. It is not my aim here to defend realism, cognitivism, or reductive ethical naturalism *per se*. Instead, my aim is to argue that Goal Theory, as a *particular* realist, cognitivist, and reductive ethical naturalist theory, resists the challenges in question.

First, error theorists would argue that Goal Theory is surely mistaken, because it is committed to there being moral facts, yet there are no such things. If they were right about this, then Goal Theory's view on the nature of moral reality would be profoundly mistaken. For example, by means of a positive conceptual/semantic claim and a negative metaphysical claim, Mackie argued (in his argument from queerness) that: (1) our concept of a moral fact is a concept of an objectively and categorically prescriptive requirement[65]; (2) there are no objectively and categorically prescriptive requirements (because such an entity, quality, or relation would be 'utterly different from anything else in the universe', being 'intrinsically action-guiding and motivating', and apprehending such requirements would then be by means of 'some special faculty of moral perception or intuition,

[65] By which Mackie means that it must be mind-independent, in the sense of existing even if people did not, and be 'intrinsically action-guiding and motivating' (irrespective of an agent's desires). Mackie, *Ethics: Inventing Right and Wrong*, p. 40.

utterly different from our ways of knowing everything else'); and so; (3) there are no moral facts.[66]

Now, the conclusion of Argument 1 from section 2.1 is compatible with the error theorist's claim that there are no moral facts, since it allows for the possibility of there being *no* true moral system. However, I subsequently argued that Goal Theory's moral facts *do* obtain (contra error theory). Let me call a moral fact on Mackie's conception an Mfact. I agree with Mackie that Mfacts do not exist, insofar as I agree that there is nothing in the world that is objective and categorical in the sense that he describes. However, Goal Theory does not conceive of moral facts in this way. Rather, it holds that our concept of a moral fact is a concept of an objectively (in the sense of being independent of people's beliefs or attitudes, but not in the stronger sense of being mind-independent) and *hypothetically* prescriptive requirement.[67] Call this distinct concept of a moral fact a GFact. In light of this conceptual distinction, then even if there are no MFacts, this, in and of itself, does not undermine Goal Theory, as there may still be GFacts. Accordingly, Mackie's argument misses its target with my account, since he and I do not agree about the features that any moral facts would have.

At this point, the error theorist might charge the Goal Theorist with conceptual confusion, and suggest that they now face a dilemma: either avoid being directly undermined by there being no MFacts, but at the cost of being conceptually confused (by endorsing GFacts), or remedy this conceptual confusion (by endorsing MFacts instead), but then have their account undermined by there being no MFacts. However, I submit it is the error theorist who is conceptually confused here, since I have already shown (in section 2.1) that Goal Theory's concept of a moral fact as an

[66] Mackie, *Ethics: Inventing Right and Wrong*, p. 39.

[67] Incidentally, John McDowell deploys a companion in innocence argument to argue against Mackie's particular mind-independence condition: McDowell, *Mind, Value, and Reality*, pp. 112-50.

objectively and hypothetically prescriptive requirement plausibly follows from an appeal to the community's linguistic intentions. I shall also argue in section 3.2 that categorical reasons are *not* a non-negotiable element of morality. Moreover, I have argued in sections 2.4 and 2.6 that making a sincere moral judgement on my account would motivate (almost) anyone to comply with it, and would generate excellent reasons for compliance — thereby plausibly capturing enough of what is insisted upon by Mackie to be, overall, plausible, without rendering moral facts too queer to posit in our overall ontology. As such, I would submit that Mackie's argument from queerness is more a reductio of his conception of a moral fact than it is a rationally persuasive argument for anti-realism.

Goal Theory might also be challenged by expressivists, who could argue that, as a realist and cognitivist account, Goal Theory is vulnerable to arguments that target these views. For example, they might leverage Blackburn's reasons to reject realism, by arguing that: (1) Goal Theory is metaphysically and epistemologically extravagant; (2) if we accept that moral judgements are internally connected to motivation, and also endorse Humean psychology (as Blackburn and I do), then it seems that moral judgements cannot be *beliefs* (in line with Smith's 'moral problem'), but instead some desire-like state (implying expressivism; but undermining cognitivism, and thence Goal Theory); and (3) that it would be mysterious on Goal Theory why the supervenience of moral facts on natural facts would ban what Blackburn calls 'mixed worlds'.[68]

On the first challenge, I would answer that Goal Theory — as a reductive naturalist account — is as ontologically parsimonious as accounts (such as expressivism) that deny an objective moral reality (positing the same natural facts and properties, and differing only from anti-realist theories in claiming that some of

[68] Blackburn, *Spreading the Word*.

these facts and properties are also referents of moral terms[69]). Goal Theory is also no more epistemologically extravagant than expressivism — locating the domain of morality within the familiar natural world, meaning that its moral facts and properties are therefore (in principle) discoverable by the familiar methods of science.

The second challenge I have already discussed in section 2.4, so I shall move on to the third. Let me first define the supervenience of the moral on the non-moral as follows: let *N* be a complete description of all of the natural facts of an act, situation, or event. In that case, if two acts, situations, or events are *N*, then they must also be morally identical. Blackburn contrasts this with the stronger notion of *necessitation*, whereby, in any possible world, all of the moral facts of an act, situation, or event are determined by *N*. Thus, for any given moral fact *M*, it is necessarily the case that if an act, situation, or event has *N*, then it has *M*. Blackburn finds necessitation of this sort implausible. As he says:

> It does not seem a matter of conceptual or logical necessity that any given total natural state of a thing gives it some particular moral property. For to tell which moral quality results from a given natural state means using standards whose correctness cannot be shown by conceptual means alone. It means moralizing, and bad people moralize badly, but need not be confused.[70]

However, notwithstanding Blackburn’s scepticism about necessitation, if Goal Theory is the true moral theory, then it is correct. As explained, on Goal Theory, moral facts are reductively identical to natural facts. Thus, for a given moral fact *M*, it is necessarily the case that if an act, event or situation has *N*, then it has *M*. That is, if it is a moral fact that agent *A* ought to do *x* in circumstances *C* (as *x* is the act that

[69] Neil Sinhababu expresses the same view regarding reductive naturalist accounts: Neil Sinhababu, 'Ethical Reductionism', <http://philpapers.org/archive/SINER.pdf>, p. 17.

[70] Blackburn, *Spreading the Word*, p. 184.

will best serve *A*'s true strongest desire in *C*), then it is necessarily the case that in any possible world with the identical natural facts (including of human psychology and cause and effect), it will be a moral fact that *A* ought to do *x* in *C*.

Blackburn's argument is then that supervenience allows the creation of worlds like the following, which contains only one individual object, *b*:

(W1): *b* is *N* and *b* is not *M*

This world is allowed since supervenience only says that if two things are *N*, then they must also be *M* — but there is only one object in (W1). However, supervenience would ban the following world, since it now contains (at least) two objects:

(W2): *a* is *N* and *a* is *M*, *c* is *N* but *c* is not *M*

However, the ban on world (W2*)* is then mysterious if (W1) is permitted, since (W2) is relevantly similar to (W1), but with the addition of *a* being *N* and *a* being *M*. As Blackburn says:

> These questions are especially hard for a realist. For he has the conception of an actual moral state of affairs, which might or might not distribute in a particular way across the naturalistic states. Supervenience [and the ban on mixed worlds] then becomes a mysterious fact, and one of which he will have no explanation (or no right to rely on). It would be as though some people are N and doing the right thing, and others are N but doing the wrong thing, but there is a ban on them travelling to the same place: completely inexplicable.[71]

[71] Blackburn, *Spreading the Word*, pp. 185-86.

However, by being committed to necessitation, Goal Theory is immune to this argument. On necessitation, if what is *N* is *M*, then world (W1) is ruled out, and so we do not get 'mixed' worlds.

Moving on, Cornell Realists might charge Goal Theory, as a *reductive* form of synthetic naturalism, with being vulnerable to multiple realizability considerations of the sort that have led to non-reductionism becoming the dominant view amongst synthetic naturalists. As Miller says of multiple realizability:

> We can imagine an indefinite number of ways in which actions can be morally right. Non-reductionist naturalistic cognitivists think that in any one example of moral rightness, the rightness can be identified with Natural properties (e.g., being the handing over of money, being the opening of a door for someone else, etc.). But they claim that across all morally right actions, there is no one Natural property or set of Natural properties that all such situations have in common and to which moral rightness can be reduced.[72]

For example, David Brink offers two particular multiple realizability challenges against reductionism, both of which may be aimed at Goal Theory.[73] Firstly, even in a world with different fundamental properties, moral properties like wrongness could be realized (so that, for example, in a nonphysical world, it would be wrong for ghosts to torture other ghosts). Secondly, in actuality, wrongness is realized by many different physical structures, so that gender discrimination, slavery, and aliens torturing other aliens have little in common at the level of physics that distinguishes them from things that are not wrong.

In answer to the first of Brink's challenges, the Goal Theorist would say that their theory is not committed to desires — as particular mental states such that one

[72] Alexander Miller, *Contemporary Metaethics: An Introduction*, 2nd edn (Cambridge: Polity Press, 2013), p. 144.
[73] Brink, *Moral Realism and the Foundation of Ethics*, p. 178.

has a disposition to act — depending on some specific internal constitution. As such, the Goal Theorist may adopt a functionalist position on what makes something a mental state of a particular type, meaning that so long as a moral agent (whether a ghost, an alien, or an intelligent robot, for example), has something that *plays the role* of a desire (regardless of the internal constitution that realizes this mental state), then, in principle, moral properties can be realized. In the case of the ghosts, for example, if they have particular mental states (realized in some nonphysical way) such that they have dispositions to act, these 'desires' can be stronger or weaker, ghosts can in principle be fully rational and sufficiently informed, and their different actions can produce results that serve their true strongest desires to greater or lesser extents, then 'rightness' (as the property that actions have when they best serve the true strongest desire of that 'person' in those circumstances) can be realized (and so on for other moral properties). Given these conditions, then, on the assumption that in their world a ghost does not best serve its true strongest desire by torturing other ghosts, on Goal Theory it would be wrong for them to do so.

With regard to Brink's second challenge, the Goal Theorist would reply that their identity of moral properties with natural ones is at the level of *psychological* properties (relating to desire satisfaction), where we get a type-reduction of moral properties to individual psychological properties.[74] This is as opposed to being a reduction at the level of *physics*, where we might get a problematic *infinite* disjunction (because psychological properties might be infinitely realizable at the level of physics).[75] Consequently, that wrongness is realized by many different

[74] However, even if we had a type-reduction of moral properties to a *finite* disjunction of psychological properties (e.g. moral property *P* is realized by x_1, or x_2, or …, x_n, where *n* is finite), this may still be unproblematic, since reductionism allows as many disjuncts as one wants.

[75] For a detailed defence of the view that reductionist accounts are not troubled by multiple realizability, so long as they do not treat moral properties as being infinitely disjunctive), see: Sinhababu, 'Ethical Reductionism'.

physical structures, with particular instances of this having little in common at the level of physics that distinguishes them from things that are not wrong, is immaterial to the Goal Theorist. At the level of Goal Theory's reduction, all instances of wrongness for an agent *do* have something in common, being whatever fails to best serve the true strongest desire of the moral agent in question. On Goal Theory, this clearly distinguishes them from things that are not wrong.

The Goal Theorist might also be asked why we should call the system of imperatives or oughts that is generated by how to fulfil our true strongest desire 'moral'. Instead, is this not more of a proposal to *eliminate* morality as a category of either action or imperative? In response, I would suggest that if I was merely identifying Goal Theory's system of imperatives as the set of imperatives determining what we ought most to do, whether we label them as 'moral' or not, then this objection might gain some traction. However, based upon the approach I have adopted here, this would be to get things back to front. Specifically, I did not begin with Goal Theory's system of imperatives, and then attempt to justify why this system should be called 'moral'. Rather, I began with a plausible definition of 'morality', went through several subsequent steps that were themselves plausible, and ended up deriving Goal Theory's system of imperatives. In so doing, my argument for claiming that what we ought to do is that which best satisfies our true strongest desire is a mix of the empirical (what we find people intend to refer to with their use of the word 'moral'), and the conceptual (what may be inferred from this).

Next, on my account, is there such a thing as *practical reason*, apart from means-end reasoning? Well, if practical reason is understood as the capacity for argument or demonstrative inference, considered in its application to the task of prescribing or selecting behaviour, then I adopt an *instrumentalist* position in terms of the role that practical reason plays in determining norms of conduct. On this position, reason fulfils an indispensable function in discerning means-end relations by which our goals may be attained. However, none of those goals is *set* by reason. Rather, all are set by our *desires*. Hence, on my view, moral injunctions must be grounded in desires, and practical reason is of interest only as subordinated to inclination. It follows from this that I reject the Kantian view that practical reason is an autonomous source of normative principles, with desires lacking intrinsic moral import, and with the function of practical reason being to limit their motivational role by formulating normative principles binding for all rational agents and founded in the operation of practical reason itself (with the moral principles in question typically grounded in consistency, and an impartial respect for the autonomy of all rational agents). This is a subject to which I shall return in the next chapter.

Can any sense be given on Goal Theory to the phrase ‘what one ought to want’? Given that we do commonly employ this idea, then, if Goal Theory denies it, can the claim be defended? In answer to the first question, I would say yes, in one sense: what one ought to want is what one would want if one were fully rational and sufficiently informed. And that is a perfectly coherent proposition (and empirically discoverable, or so I would argue). Moreover, for reasons to do with our shared fundamental biology and environment, I would suggest that what we all ought to want will mostly align in terms of our more fundamental wants (e.g. health, pleasure, achievement, love, and friendship). However, I allow for the possibility of exceptions to this; and, in any such cases, I think it would then *not* make sense to say that people ‘ought’ to want something that they would not want if they were fully rational and sufficiently informed, merely because most other people would want that thing.

Finally, I come to three challenges (two metaethical and one normative) to Goal Theory's commitments that are so dominant I think each warrants a much more detailed critical evaluation. Firstly, that there are *categorical* normative reasons, contra Goal Theory's Humean account. This widely held notion of morality's categorical reason-giving power underdetermines one's account of the nature of moral reality. As such, we find, for example, that Richard Joyce, Michael Smith, and Russ Shafer-Landau all endorse it, despite them holding radically different metaethical views.[76] According to the next objection, normative facts and properties are *just too different* from natural facts and properties to be reducible or identical to them, contra Goal Theory's naturalistic account of normativity. David Enoch identifies this 'just too different' intuition as the underlying motivation for non-naturalism.[77] (One might also be drawn to non-naturalism — thereby rejecting Goal Theory — if it was found that an Open Question Argument could also be formulated to challenge *non*-analytic naturalism, despite non-analytic naturalism rejecting the claim that property identity requires synonymy. This is something that I shall address in section 6.4.) On the last objection, ethical egoism succumbs to a number of internal and external criticisms, thereby undermining Goal Theory's egoist account.[78]

If Goal Theory cannot plausibly resist these three challenges, then this would constitute a serious strike against the theory.

[76] See, for example: Cuneo, 'Moral Naturalism and Categorical Reasons'; R. Shafer-Landau, 'A Defence of Categorical Reasons', *Proceedings of the Aristotelian Society (Supp.),* 109 (2009), 189-206.

[77] David Enoch, *Taking Morality Seriously* (Oxford: Oxford University Press, 2011).

[78] See: James Rachels, 'Ethical Egoism', in *Ethical Theory: An Anthology,* ed. by Russ Shafer-Landau (Hoboken, NJ: John Wiley and Sons, 2012), pp. 193-99.

2.8 Conclusions

I began this chapter by formulating a positive argument for Goal Theory — finding that, if there is a true moral system at all, then it is probably Goal Theory. As a first-order theory, I found Goal Theory to be a variety of ethical egoism, with it conceiving of morality as a system of hypothetical imperatives. Having done that, I unpacked and examined Goal Theory's commitments along various dimensions of metaethics. Here, I found it to be a realist, cognitivist, and reductive naturalist account, with its moral facts being reductively identified with complex natural facts composed of idealised (i.e. fully rational and sufficiently informed) human desire and cause and effect. I further found that its moral facts and properties are in principle discoverable by the familiar methods of science, that it is able to offer a plausible explanation of why it is that (almost) anyone who makes a sincere moral judgement on the theory would be motivated to some extent to comply with it, and that it explains why agents will have excellent reasons for compliance with moral requirements. I then suggested that its moral facts are probably (almost) universal, ventured that they probably align with our stock moral truisms, and proposed a plausible candidate for a universal true strongest desire. Finally, I considered some possible questions and objections. Most were dealt with swiftly, but three of these objections I deemed to be substantial enough to warrant a more detailed treatment in subsequent chapters.

Accordingly, I think I have provisionally established Goal Theory as a plausible candidate for the true moral system postulated earlier — with it providing credible answers to questions about the nature of moral reality (explaining in particular why there are objective moral facts and properties, and how these fit within the natural world, with no need to add any *sui generis* facts and properties to our ontology), in addition to questions about the possibility of moral knowledge (with these facts and properties being in principle discoverable by the familiar methods of science, rather than by appeal to some special faculty or other means by which we apprehend non-natural *sui generis* facts and properties), questions about the connection between moral thought and motivation and reasons for action, and questions about the connection between morality and self-interest.

Of course, more work is required to strengthen these answers, with them being merely tentative at this stage. To that end, I have also laid the foundations for what is to come in the next four chapters, where I intend to demonstrate that Goal Theory probably resists the three objections mentioned above, in addition to meeting the applicable theoretical adequacy criteria described in section 1.1. The first of these challenges — i.e. that there are *categorical* normative reasons, contra what Goal Theory would claim — will be the subject of the next chapter.

Chapter 3

The challenge from the existence of categorical reasons

Why ought we to tell the truth? Why should we be compassionate? More generally, why should we be moral at all? Whatever the answers, is it not evident to any conceptually competent person that the reasons must apply to us regardless of our contingent desires? Surely that is an essential element of morality, with its absence potentially supplying people with a normative reason to lie, cheat, steal, or murder, so long as these immoral actions would serve their desires, but performing alternative moral actions would not. Accordingly, is it not manifest that there are *categorical* normative reasons?

In asking questions like these, we are seeking an account of the rational authority of morality, as the subject of whether or not moral requirements supply good reasons for obedience. This falls within the general area of *normativity*, where this may be broadly understood as the property of expressing or essentially concerning *reasons* — to act, to believe, to feel, or to want, for example.[1] Here, I shall focus upon the first of these; and so, for this purpose, my concept of a normative reason will be a concept of a reason to act.

What kinds of reasons might there be? There may be many, but all parties to the debate would agree that there are *hypothetical* reasons, as reasons that derive from their

[1] For example, Derek Parfit thinks that ‘normativity is best understood as involving reasons or apparent reasons.’ Parfit, *On What Matters Vol 2*, p. 269.

relation to agents' desires (or similarly conative attitudes — but henceforth I shall focus upon desires). After all, if we want to stay dry outside on a rainy day, then we have a *reason* to take an umbrella with us. Likewise, if we want to do well in an exam, then we have a reason to study; and if we want to be treated well by others, then we have reason to treat them well. None of these reasons is necessarily decisive, insofar as each may in principle be outweighed by reasons to act otherwise, but they are reasons nonetheless. However, some want to go further than this, by postulating a kind of reason that obtains independently from its relation to agents' desires. If such *categorical* reasons exist, then, even if an agent has no desire that would be served by doing such and such, they may still have a reason to do it.[2]

Of particular relevance here, the proponent of categorical reasons would want to claim that at least some *moral* requirements generate categorical reasons, such that we have excellent reasons to do those things, independently of any desire that we may possess. However, for those who would deny categorical reasons, if an agent has any reason to act as morality demands, then such a reason will be a hypothetical one, contingent upon serving the desires of the agent concerned.

In this chapter, I intend to answer the challenge that there are categorical normative reasons, contra Goal Theory's Humean account. To that end, I shall begin by evaluating the resistance of the HTR* to the so-called Central Problem, as well as to undergeneration and overgeneration arguments (where these are often held to be problematic for the HTR). Next, I shall critically evaluate two representative arguments for categorical reasons — one from Richard Joyce and the other from Russ Shafer-Landau. My aim is to show that Goal Theory survives both. I shall then ask who bears the burden of proof here.

[2] As I shall understand them, categorical reasons need not of necessity override all competing reasons, nor necessarily apply to all rational agents (though some would endorse these additional claims).

3.1 The Central Problem

Before I examine Joyce’s and Shafer-Landau’s specific arguments for categorical reasons, I want first to address a general set of challenges facing the Humean theory of reasons, showing that the HTR* (on which an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent) plausibly resists them. Within the context of the chapter, this means that I can deny categorical reasons without falling into an objectionable theory.

Although many moral naturalists endorse the HTR (i.e. an agent has *pro tanto* normative reason for an action just in case that action would serve some of the agent’s desires), it is not uncontroversial. There are significant objections to the theory, with these coalescing chiefly around what is known as the *Central Problem*, whereby there is a tension between the HTR, *Moral Rationalism* (according to which, if something is morally wrong then there must be a reason not to do it), and *Moral Absolutism* (according to which, some actions are morally right or wrong for any agent, no matter what desires they possess), such that we cannot consistently hold all three views simultaneously. Specifically, if an action is morally wrong for an agent just in case there is a reason for them not to do it (as Moral Rationalism claims), and if there is a reason for them not to do it just in case they have some desire that would be served by them not doing it (as HTR claims), then it follows that whether an action is morally wrong for an agent depends upon what they desire. However, that then appears incompatible with Moral Absolutism.

For example, if, as Moral Rationalism says, ordering genocide was morally wrong for Hitler just in case there was reason for him not to do it; and if, as the HTR says, there was reason for him not to do it just in case he had some desire that would be served by him not doing it; then whether ordering genocide was morally wrong for Hitler would depend upon his desires. However, this would be incompatible with a Moral Absolutist view that ordering genocide was morally wrong for Hitler no matter what his desires.

In light of this tension, some philosophers reject Moral Absolutism (e.g. Harman, Mackie, and Joyce[3]), and others reject Moral Rationalism (e.g. Foot[4]). However, some others prefer to keep the *prima facie* commonsense moral views expressed by Moral Rationalism and Moral Absolutism, and to reject the HTR instead (e.g. Dancy, Raz, Scanlon, and Korsgaard[5]). More generally, some would argue the HTR *undergenerates* reasons (failing, in particular, to generate moral reasons in some instances where we intuitively believe there are such reasons), while simultaneously *overgenerating* reasons (in counting some things as reason-giving that clearly are not).

I shall set aside questions relating to how a proponent of the HTR might respond to the challenge of the Central Problem, since I do not endorse the HTR, but instead the HTR*. So, how does the HTR* fare against the Central Problem, as well as against undergeneration and overgeneration objections? With regard to the Central Problem, if an action is morally wrong for an agent just in case there is a reason for them not to do it (as Moral Rationalism claims), and if there is a reason for them not to do it just in case a fully rational and sufficiently informed version of themselves would have some desire that would be served by them not doing it (as HTR* claims), then it follows that whether an action is morally wrong for an agent depends upon what they would desire, if they were fully rational and sufficiently informed (call these

[3] Gilbert Harman, 'Moral Relativism Defended', *Philosophical Review*, 85 (1975), 3-22; Richard Joyce, *The Myth of Morality* (Cambridge: Cambridge University Press, 2001); Mackie, *Ethics: Inventing Right and Wrong*.

[4] Foot, 'Morality as a System of Hypothetical Imperatives'.

[5] J. Dancy, *Practical Reality* (Oxford: Oxford University Press, 2000); Christine Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996); J. Raz, *Engaging Reason: On the Theory of Value and Action* (Oxford: Oxford University Press, 1999); T. M. Scanlon, *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998).

'enlightened' desires, with desires then being 'unenlightened' if they are not enlightened). However, that then appears incompatible with Moral Absolutism. Therefore, it seems that there is a tension between the HTR*, Moral Rationalism, and Moral Absolutism, implying that we must reject at least one of them. I would reject the last of these, since Goal Theory is incompatible with Moral Absolutism. Specifically, Goal Theory's conception of moral rightness is based upon agents' (true strongest) desires, and so there can be no actions that are morally right or wrong for any agent no matter what desires they possess (as Moral Absolutism claims).

By contrast, Goal Theory is compatible with the other two members of the triad. Firstly, I have already shown (in section 2.6) that the HTR* follows directly from a conjunction of Goal Theory, Proportionalism, and Parfit's platitudes. As for Moral Rationalism, we may say that, on Goal Theory, action *x* will be morally wrong for agent *A* in circumstances *C* just in case doing *x* in *C* will not best serve the strongest desire of a fully rational and sufficiently informed version of *A*. In that case, doing some particular ~*x* *would* best serve the strongest desire of a fully rational and sufficiently informed version of *A* in circumstances *C* (as there must be *some* action in *C* that will best serve the desire in question, and, *ex hypothesi*, it is not *x*). But if doing some ~*x* would best serve the strongest desire of a fully rational and sufficiently informed version of *A*, then, on the HTR*, *A* then has a *reason* to not do the morally wrong action *x*. Accordingly, an action is morally wrong for an agent on Goal Theory just in case there is a reason for them not to do it — as Moral Rationalism requires.

That is not the end of the story though, because I would now like to propose a weaker version of the moral absolutism claim that is *not* in tension with the HTR* and Moral Rationalism, and with which Goal Theory is plausibly compatible:

> **Moral Absolutism***: some actions are morally wrong for any agent, no matter what *unenlightened* (e.g. present) desires they possess.

As before, it follows on the conjunction of Moral Rationalism and the HTR* that whether an action is morally wrong for an agent depends upon what they would desire, when fully rational and sufficiently informed (i.e. their enlightened desires). However, this does *no*t then appear incompatible with Moral Absolutism*, since moral wrongness not depending upon unenlightened desires is compatible with it depending upon enlightened ones (though the incompatibility would return if we replaced the HTR* with the HTR). Thus, there would appear to be no tension between the HTR*, Moral Rationalism, and Moral Absolutism* (thereby resisting the Central Problem).

As for Goal Theory's compatibility with Moral Absolutism*, observe that there will be no conflict between the two if either of the following obtains: (1) all agents share the *same* true strongest desire, and the *same* actions in certain moral circumstances will or will not best serve this desire; or (2) every agent has a true strongest desire (not necessarily the same one) that will or will not be best served by the same actions in certain moral circumstances. Here I would endorse the former option, noting that I have already defended it (on the grounds of shared fundamental biology and environment) in section 2.5. As such, we might find, for example, that committing genocide is morally wrong on Goal Theory for any agent, no matter what unenlightened (e.g. present) desires they possess, because all agents have the same true strongest desire (e.g. a deep and abiding satisfaction), and committing genocide will not best serve this true strongest desire for any agent.

Let me now revisit the Hitler example discussed earlier. If, as Moral Rationalism says, ordering genocide was morally wrong for Hitler just in case there was reason for him not to do it, and if, as the HTR* says, there was reason for Hitler not to do it just in case not doing it would serve some of the desires of a fully rational and sufficiently informed version of himself, then whether ordering genocide was morally wrong for Hitler would depend upon his *enlightened* desires. However, this is then not incompatible with a Moral Absolutist* view that ordering genocide was morally wrong for Hitler no matter what his *unenlightened* desires. Moreover, I suggest his enlightened desires would almost certainly be such as to render it wrong for him to order genocide.

Accordingly, I would argue that my account captures what I think is the underlying intuition behind Moral Absolutism, viz. that what is morally wrong for an agent cannot be dependent upon their *unenlightened* (e.g. present) desires (or else Hitler would not have been morally wrong to order genocide, if this served some of his present desires, without being detrimental to any of them). Yet it denies what I think is the implausible claim that some actions are morally right or wrong for any agent, no matter what *enlightened* desires they possess. At the same time, it does not endorse what I think is the equally implausible view that the possession of any desire, no matter how unenlightened, yields a reason for an agent to act accordingly (as the HTR declares). In light of this, I would suggest that my account occupies a plausible position between implausible extremes.

As for the undergeneration objection, the claim would be that the HTR* will produce *too few* reasons. In particular, we intuitively believe that there are some paradigmatic right or wrong actions that anyone has a reason to do or not do. However, by making normative reasons contingent upon desires, the HTR* might fail to generate these reasons in some instances. In defending *hypotheticalism* (i.e. his proposed variant of the HTR, on which, necessarily, an agent has a reason to do *X* if and only if doing *X* will promote one of the agent's desires) against the same charge, Schroeder relies upon a massive overdetermination of agent-neutral reasons, claiming that acting rightly always promotes at least one of an agent's desires, regardless of what those desires are. Whether he is successful is a moot point[6], but I shall adopt a different approach.

I would argue that for reasons to do with our shared fundamental biology, conscious experience, and environment: (1) (almost) all agents' share the same true strongest desire, viz. a deep and abiding satisfaction; and (2) doing the kind of paradigmatic right actions that we intuitively believe anyone has a reason to do will serve this desire for (almost) all agents in (almost) all circumstances, whilst doing paradigmatic wrong actions will frustrate this desire for (almost) all agents in (almost) all circumstances. (Both of these claims were defended in section 2.5). In particular, I would suggest that acting cooperatively, compassionately, altruistically, and honestly would probably serve the desire for deep and abiding satisfaction for (almost) all agents in (almost) all circumstances; whereas acting uncooperatively, callously, selfishly, or dishonestly will frustrate this desire.

Accordingly, I suggest the HTR* will very probably generate *pro tanto* normative reasons in (almost) all paradigmatic cases where we intuitively believe there really are such reasons (e.g. to rescue a child in imminent danger, to avoid killing another person simply because it gives one pleasure, and to avoid breaking a promise just because one feels like it), since acting in these ways will very likely serve the strongest desires of fully rational and sufficiently informed agents. The reasons in question might not always be *decisive* ones (after all, one might have to risk one's life to save the child in danger, and the reason for one to avoid this risk might in principle outweigh the reason to rescue the child), but it is not part of the requirement that the reasons be decisive.

[6] Shafer-Landau for one thinks not: R. Shafer-Landau, 'Review: Three Problems for Schroeder's Hypotheticalism', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* 157 (2012), 435-43.

The next challenge is to avoid generating *too many* reasons, by counting things as reason-giving that are not. In accord with Shafer-Landau's criticism of Schroeder's hypotheticalism, we may say that, firstly, we would not want to generate reasons based upon uninformed desires.[7] Shafer-Landau argues that Schroeder's account is vulnerable to this objection. However, in claiming that an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and *sufficiently informed* version of the agent, the HTR* explicitly excludes the generation of normative reasons based upon uninformed desires.

Secondly, we would want to exclude reasons based upon desires that are for worthless things, and where the satisfaction of the desire is valueless or positively disvaluable.[8] Again, Shafer-Landau charges Schroeder's account with being vulnerable to this objection. However, in basing normative reasons for an agent upon the desires of a fully rational and sufficiently informed version of the agent, I would submit that my account also avoids this unhappy outcome, on the basis that such an ideal agent would have epistemic access to what is (independently) valuable and would not hold desires for things that are not.

In accord with McPherson's analysis, I would also want my account to avoid the problematic *explosion* of agent-neutral reasons that Schroeder's hypotheticalism is charged with.[9] Remember, on the HTR*, an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. I have not yet specified under what conditions an action could be said to *serve* a desire, so I shall remedy that now. On my conception, this *serving*-relation is analogous to Schroeder's promotion-relation.[10]

[7] Shafer-Landau, 'Review: Three Problems for Schroeder's Hypotheticalism', p. 436.
[8] Shafer-Landau, 'Review: Three Problems for Schroeder's Hypotheticalism', pp. 436-37.
[9] Tristram McPherson, 'Review: Mark Schroeder's Hypotheticalism: Agent-Neutrality, Moral Epistemology, and Methodology', ibid., 445-53.
[10] Schroeder, *Slaves of the Passions*, p. 113.

Specifically, I would claim that *A*’s doing *x serves* desire *d* just in case it increases the likelihood of *d* being satisfied relative to some baseline, with the baseline being fixed by the probability of *d* being satisfied conditional on *A*’s doing nothing (i.e. conditional on the status quo).

In requiring only that the likelihood be increased relative to the status quo, this serving-relation might be said to be a *relaxed* one. Schroeder needs such a relaxed notion of the serving-relation in order to disarm the too few reasons objection against hypotheticalism. In my case, however, I do not require it for this purpose. Instead, I endorse it merely because I think it is plausible. Specifically, on the conjunction of this serving-relation and the HTR*, the following (plausible) claim is implied:

> **(S)**: an agent has a *pro tanto* reason for an action just in case doing that action would increase the likelihood of satisfying one of the desires of a rational and sufficiently informed version of the agent, relative to the likelihood of satisfying one of these desires if the agent does nothing.

So, with this understanding of the serving-relation in place, does my account suffer from an explosion of agent-neutral reasons, in the way that Schroeder’s account arguably does? No — my account stops short of generating McPherson’s explosion, because there will not be agent-neutral reasons to do everything. In particular, and as already explained, because the desires being served are those of fully rational and sufficiently informed versions of the agents, rather than of the ordinary agents themselves, then it will not yield uninformed reasons, or reasons based upon desires that are for worthless things, or reasons where the satisfaction of the desire is valueless or positively disvaluable. Moreover, when I factor in my weighting scheme for reasons (i.e. Proportionalism, on which, when a reason is explained by a desire, its weight varies in proportion to the strength of that desire, and to how well the action promotes that desire), then my account allows us (in principle) to assign different weights to reasons, with some being assigned relatively low weights. For practical purposes, we might then want to introduce a weight threshold, below which any reasons would effectively be disregarded.

To see how this might work in practice, consider an example discussed by Schroeder, viz. having a reason to eat one's car if one needs to get the requisite dose of iron in one's diet, and eating one's car would satisfy this requirement.[11] In this case, the desire for iron is generally a relatively weak one. More importantly, eating a car is a very poor way of promoting that desire, since it is both highly impractical and hazardous to health, and there are far safer and more practicable alternatives available. As such, although my account would generate a reason to eat one's car under the conditions described, I suggest the reason's weight would fall below any reasonable weight threshold, meaning that it may effectively be disregarded.

Accordingly, I think my account avoids the problematic (and potentially fatal) worry to which Schroeder's hypotheticalism is arguably vulnerable, whereby we have a superabundance of reasons (including reasons for uninformed, worthless, or positively disvaluable things), with no reasons being weightier than any others.

In conclusion, I submit that the HTR* offers a *prima facie* plausible solution to the Central Problem, in addition to credibly resisting the undergeneration and overgeneration objections (neither undergenerating nor overgenerating reasons). In this regard, I suggest that it improves upon other versions of the HTR, including Schroeder's hypotheticalism. Now, having established that I can deny categorical reasons without falling into an objectionable theory, I return to the primary purpose of

[11] Schroeder, *Slaves of the Passions*, p. 95.

this chapter, viz. to critically evaluate the objection that there are categorical normative reasons, contra Goal Theory's Humean account.

3.2 Joyce's argument

Richard Joyce endorses moral error theory, which holds that our (positive, atomic) moral judgements are truth-apt, but uniformly and systematically false, because there simply are no moral facts or properties in the world of the sort required to render our moral judgements true. However, as a revolutionary fictionalist, Joyce's response to this supposed uniform and systematic moral error is that we ought to engage in a convenient fiction — make-believing that certain moral claims are true — on the basis that there are practical benefits to doing this (e.g. in terms of facilitating agreement and coordination).[12]

In this chapter, I am not engaging with error theory *per se*, so why am I invoking Joyce? Well, according to Joyce, there is an element of morality that is conceptually *non-negotiable*, viz. its categorical reason-giving power.[13] In that case, for a moral naturalist account (such as Goal Theory) to be a tenable one, it must accommodate the proposition that, necessarily, if someone morally ought to do something, then they have *categorical* reason to do that thing; with the denial of this categoricity claim (if widely believed) disabling some crucial uses of moral concepts (e.g. the concepts of moral obligations and prescriptions, as well as moral rightness and wrongness). Joyce goes on to infer his error theoretic conclusion by arguing that there are *no* such categorical reasons. Clearly, it would be beneficial to me if this last

[12] This is as opposed to adopting an *eliminativist* position, by giving up moral discourse altogether. See: Joyce, *The Myth of Morality*; Richard Joyce, *The Evolution of Morality* (Cambridge, MA: MIT Press, 2006).

[13] Joyce, *The Myth of Morality*, pp. 3-4, 176-77.

argument were cogent, since this would then provide me with a direct response to the main challenge of the chapter. Joyce does not lay out his argument against categorical reasons syllogistically, but it may be formulated as follows:

> (1) If there are categorical reasons, then, for any reasonable agent *S*, *S* might have reason to Φ, but fail to be engaged by Φ.
>
> (2) Reasons cannot fail in this way.
>
> (3) Therefore, there are no categorical reasons.[14]

Unfortunately, I find the argument unpersuasive, agreeing with Shafer-Landau that, in an argument designed to impugn categorical reasons, premise (2) is question-begging as it stands.[15] However, Joyce does not defend that premise, beyond claiming that, necessarily, reasons must be capable of *engaging* (e.g. motivating, sparking interest, providing an affirmative answer to the question of a consideration's importance) the agent whose reasons they are. Given this, and because I can see no obvious way to repair the argument, then I shall set aside any attempt to press the case against categorical reasons by means of such a direct argumentative strategy.

Joyce's core argument for his error-theoretic conclusion may be reconstructed as follows:

> (1) Categorical reasons are a non-negotiable element of morality.
>
> (2) There are no categorical reasons.
>
> (3) Therefore, morality is a fiction.

[14] See: Joyce, *The Myth of Morality*, pp. 39-45.

[15] Russ Shafer-Landau, 'Error Theory and the Possibility of Normative Ethics', *Philosophical Issues*, (2005), 107-20 (p. 114).

Rearranging this argument, we may formulate an argument for categorical reasons:

Argument 2

P1)	If categorical reasons are a non-negotiable element of morality, and if morality is not a fiction, then there are categorical reasons.
P2)	Categorical reasons are a non-negotiable element of morality.
P3)	Morality is not a fiction.
C)	Therefore, there are categorical reasons.

This argument is valid, so because I want to deny the conclusion, I must deny either that morality is not a fiction, or that categorical reasons are a non-negotiable element of morality. Since I am defending Goal Theory's realist account, I can hardly do the former, so my remaining option is to deny that categorical reasons are a non-negotiable element of morality. To that end, I shall critically evaluate Joyce's justification for his non-negotiability claim, arguing that it fails to do the necessary work required of it.

In order to argue for this non-negotiability claim, Joyce adduces two further claims about morality: (1) that it is *inescapable*; and (2) that it is *authoritative*. According to the former, if there is a moral obligation to do something, then this obligation is categorically applicable, in the sense that an agent is morally obligated to do that thing regardless of whether doing so serves any of their desires. And according to the latter, necessarily, if an agent morally ought to do something, then the agent has *pro tanto* normative reason to do that thing. The conjunction of the putative authority and inescapability of morality Joyce calls 'practical clout', with this supposedly entailing that there are categorical reasons to do as one morally ought to. More formally, Joyce's argument runs as follows:

Argument 3

P1)	If *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether he cares to, regardless of whether Φing satisfies any of his desires or furthers his interests. [inescapability]
P2)	If *x* morally ought to Φ, then *x* has a reason for Φing. [authority]
C1)	Therefore, if *x* morally ought to Φ, then *x* has a reason for Φing regardless of whether Φing serves his desires or furthers his interests. [categorical reasons]
P3)	But there is no sense to be made of such reasons.
C2)	Therefore, *x* is never under a moral obligation. [error theory]

Here, the sub-argument from P1-C1 represents Joyce's argument for categorical reasons being a non-negotiable element of morality, with the sub-argument from P3-C2 inferring his error-theoretic conclusion.[16] Of course, this just pushes the problem back a step, since Joyce is now required to justify the inescapability and authority claims in premises P1 and P2. In *The Myth of Morality*, he offers the following in support of inescapability:

> When we morally condemn a criminal we do not first ascertain certain the state of his desires. Were we to discover that his desires were well-served by his crimes, perhaps

[16] Joyce, *The Myth of Morality*, p. 42. This argument is slightly revised on p. 77 in order to make some concessions to the non-Humean, but is left basically intact.

> even to the point of his wanting punishment, we do not respond 'Oh, well I suppose you ought to have done it after all.'[17]

And:

> The manner in which we condemn Nazis, ignoring any unusual desires or interests that they may have, is not a peripheral element of moral discourse; it presents a kind of reprehension that is central. A system of values in which there was no place for condemning Nazi actions simply would not count as a moral system.[18]

In terms of authority, Joyce's conception captures what he calls *Mackie's platitude*, according to which it is necessary and *a priori* that, for any agent *x*, if *x* ought to do Φ, then *x* has a reason to Φ.[19] Mackie's platitude is supposed to span all normative domains, so Joyce applies it here to *moral* prescriptions, with a moral ought then entailing a moral reason.

In *The Evolution of Morality*, Joyce expands upon the aforementioned defences, with an extended justification of practical clout, arguing that it is a core desideratum of any moral theory. Moreover, he argues that no form of moral naturalism can satisfy this desideratum. As part of his argument for these claims, Joyce introduces us to the so-called 'sensible knave', whom he understands as a person for whom an immoral action will serve their desires, but performing an alternative moral action will not.[20] Joyce argues that, intuitively, even knavish people have normative reasons to act morally.

Now imagine, as Joyce does, that Jack really, really wants to murder John, and finds himself in circumstances where he could do this with very little likelihood of being caught (and so we suppose that murdering John will further Jack's desires, but

[17] Joyce, *The Myth of Morality*, pp. 42-43.
[18] Joyce, *The Myth of Morality*, p. 43.
[19] Joyce, *The Myth of Morality*, p. 38.
[20] Hume says that the sensible knave is a person for whom 'an act of iniquity or infidelity will make a considerable addition to his fortune…' David Hume, 'An Enquiry Concerning the Principles of Morals', in *Enquiries Concerning Human Understanding and Concerning the Principles of Morals,* ed. by L.A. Selby-Bigge and P.H. Nidditch (Oxford: Clarendon, 1975 [1777]), (p. 282). Railton also refers to 'knavish desire': Railton, 'Moral Realism', p. 203.

not doing so will not).[21] Why should Jack not do so? Joyce thinks the moral naturalist will agree that it appears that Jack has a moral obligation not to murder John. Yet, if morality lacked practical clout, then we must allow that any moral obligation need not necessarily generate a corresponding moral ought for Jack, and any moral ought need not necessarily generate a corresponding normative reason for Jack (since, if morality lacked inescapability, then any moral obligation to not murder John may not apply to Jack, given his desire profile; and if it lacked authority, then even if Jack has a moral obligation not to murder John, he may not have a normative reason to not do so). In fact, since, *ex hypothesi*, Jack is a knave, and murdering John will further his desires, but not doing so will not, then Jack would seem on the face of it to have reason to murder John, and no real reason to avoid doing so. In general, we may say that if morality was connected only contingently with people's reasons, then agents who possess or acquire knavish desires need not necessarily do what they *morally* ought to do.

According to Joyce, a crucial use of moral concepts is to evaluate actions, in terms of judging whether an action ought or ought not to be done, saying, for example, that '[t]he whole point of a moral discourse is to evaluate actions and persons with a particular force…'[22] Yet, if morality was connected only contingently with people's reasons (as it would be if it lacked practical clout), then it would seem to be legitimate for people with knavish desires to behave in systematically immoral ways. Moreover, if this became widely known, then the use of moral concepts to evaluate people's actions would be weakened. With regard to Jack and John, for example, Joyce says of a version of moral naturalism without practical clout:

[21] Joyce, *The Evolution of Morality*, pp. 203-05.

[22] Joyce, *The Myth of Morality*, p. x.

> If this version of moral naturalism…were correct and we were allowed to acknowledge this fact (i.e., if it were transparently correct), then there should be nothing wrong with our moral pronouncements reflecting this. Instead of ‘Killing John was unacceptable,’ we should be allowed to say ‘From the moral point of view, killing John was unacceptable.’ And to the observation that Jack’s action was wicked we should be permitted to add ‘but Jack had every reason to act wickedly on this occasion, and no real reason to refrain.’[23]

However, Joyce thinks that adding such qualifiers to our moral deliberations adds an extremely odd flavour to morality. As he says:

> We generally will not be comfortable saying that Jack’s actions were depraved and morally unacceptable and in the next breath asserting that he had no reason to refrain and that in fact committing the murder was what he ought, all things considered, to have done.[24]

Allowing that knavish individuals may have no normative reason to refrain from acting in morally unacceptable ways would cast doubt upon the firmly held belief that even knavish people ought to act in accord with moral evaluations. In that case, it would be an open question as to whether people have the necessary desires required to ensure that they have normative reasons to act morally, and potential appraisers would have to consider the possibility that moral transgressors had no reason to act otherwise, in light of their knavish desires. What is more, people might take advantage of what we might call *self-induced* knavery — such that whenever the benefits to doing so were sufficient, they might just adjust their desires in order to eliminate any normative reasons to act morally.[25]

All in all, Joyce thinks that if agents who possess or acquire knavish desires need not necessarily do what they morally ought to do (and this was widely known),

[23] Joyce, *The Evolution of Morality*, p. 204.
[24] Joyce, *The Evolution of Morality*, pp. 203-04.
[25] Joyce, *The Evolution of Morality*, p. 206.

then this would throw into doubt the conviction that those who possess or acquire knavish desires ought to act in accordance with *moral* evaluations. And if the conviction that those who possess or acquire knavish desires ought to act in accordance with moral evaluations is thrown into doubt, then moral concepts could not be employed to evaluate people's actions no matter what their desire profile.

Joyce thinks that categorical imperatives are so deeply entrenched in our moral discourse that no account that fails to incorporate them could even qualify as a moral one.[26] In an analysis of moral concepts, Joyce utilises a method devised by David Lewis.[27] According to Joyce, if we want to find out whether some *x* exists, we should start by constructing 'a list of platitudinous desiderata' describing all of the properties that *x* is believed to possess. We can then express these platitudes as a list of sentences, viz. '*x* is P_1,' '*x* is P_2,' '*x* is P_3,' and so on. The resulting sentences are then conjoined into an existentially quantified Ramsey sentence of the form '$\exists x(x$ is P_1 & *x* is P_2 & … *x* is $P_n)$'. This Ramsey sentence is then a representation of speakers' concept of *x*. The question then is whether anything exists that corresponds to this concept of *x*. If it does, then the conclusion would be that there is something that exists that has the properties people believe *x* to have. Otherwise, we should conclude that *x* does not exist, and that speakers' beliefs about the existence of *x* were mistaken. Of course, it may be that some of our beliefs about the properties that *x* possesses are in error, in which case *x* may exist even if nothing exists that corresponds to the representation of *x* (and vice versa).

In the case in question, Joyce constructs a representation of moral concepts, and asks whether any existing properties in the world might be the referents of such moral concepts. In order to construct this representation, he uses not only 'platitudinous

[26] Joyce, *The Myth of Morality*, pp. 176-77.

[27] Joyce, *The Evolution of Morality*; Richard Joyce, 'Metaethical Pluralism: How Both Moral Naturalism and Moral Skepticism May Be Permissible Positions', in *Ethical Naturalism: Current Debates*, ed. by Susana Nuccetelli and Gary Seay (Cambridge: Cambridge University Press, 2012), pp. 89-109. See also: David Lewis, 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society (Supp.),* 63 (1989), 113-37.

desiderata', but also claims about how moral concepts are characteristically used. According to Joyce, if a concept can be used in the ways that moral concepts are characteristically used, then the concept in question is a moral concept.[28] Otherwise the concept is not a moral concept (although Joyce does note a possible indeterminacy here, such that there may be a reasonable debate between the ethical naturalist and the moral sceptic as to whether an imperfect claimant, Φ*, to the concept, Φ, is 'close enough' to count as a revision of Φ, even though it cannot be put to all of the uses to which Φ can be put). In terms of a concept's ability to perform the characteristic uses of moral concepts, Joyce thinks that a concept cannot perform these characteristic uses if it would be abandoned by the community that has been using the concept for those uses. In particular, Joyce thinks that if morality could not be used to evaluate the actions of agents no matter what their desire profile, then the community that had hitherto employed the concept for those uses would abandon that concept.[29]

From this kind of analysis, Joyce concludes that a concept of morality without practical clout (and thence categorical reasons for action) is not a moral concept, and so morality (if it exists) must incorporate this feature.[30]

Joyce does not formulate his defence of inescapability syllogistically, but it may be reconstructed as follows:

[28] Joyce, 'Metaethical Pluralism: How Both Moral Naturalism and Moral Skepticism May Be Permissible Positions', p. 95.

[29] Joyce, *The Evolution of Morality*, p. 202.

[30] Of course, Joyce thinks that morality does not in fact exist, because he thinks there are no categorical reasons.

Argument 4

P1)	If morality lacked the feature of inescapability, then it would not necessarily be the case that if there is a moral obligation to do Φ, then agents who possess or acquire knavish desires [i.e. knaves] morally ought to do Φ.
P2)	If it would not necessarily be the case that if there is a moral obligation to do Φ, then agents who possess or acquire knavish desires morally ought to do Φ (and this was widely known), then this would throw into doubt the conviction that those who possess or acquire knavish desires ought to act in accordance with moral evaluations.
P3)	If the conviction that those who possess or acquire knavish desires ought to act in accordance with moral evaluations is thrown into doubt, then moral concepts could not be employed to evaluate people's actions no matter what their desire profile.
P4)	If a concept cannot be used in the ways that moral concepts are characteristically used, then the concept in question is not a genuine moral concept.
P5)	A moral concept cannot (or would not be able to) perform its characteristic uses if it is (or would be) abandoned by the community that had hitherto employed the concept for those uses.
P6)	If a moral concept could not be used to evaluate the actions of agents, no matter what their desire profile, then the community that had hitherto employed the moral concept for those uses would abandon it.

C1)	Therefore, if morality lacked the feature of inescapability, then putative moral concepts would not be genuine ones.
C2)	Therefore, morality is inescapable.

How might one respond to this argument? Should one concur with Joyce that morality is inescapable? If so, what implications does that have for my account?

There are several premises in Argument 4 that one might challenge. For example, one might argue, as Jon Tresan does, that people would continue to make moral evaluations of others, even in the face of knavish agents who lack desires to act morally (contra premise P6).[31] However, I shall focus instead upon premise P1. Let me begin by parsing the proposition in that premise, viz. if morality lacked the feature of inescapability, then it would not necessarily be the case that if there is a moral obligation to do Φ, then agents who possess or acquire knavish desires [i.e. knaves] morally ought to do Φ. The truth of this proposition turns upon: (1) what a *knave* is; (2) what *inescapability* means; and (3) our choice of moral theory. With regard to the first of these, I have previously defined a knave as a person for whom an immoral action will further his or her desires, but performing an alternative morally permitted or obligatory action would not. Note that Joyce's conception of the sensible knave leaves unspecified whether the desires being referenced are the knave's *enlightened* desires (i.e. those desires that would be possessed by a fully rational and sufficiently informed version of the knave, which would include their true strongest desire), or their *unenlightened* ones (which may include their present desires). (To suggest the conjunction of the two would be incoherent, insofar as what furthers one's

[31] Jon Tresan, 'Question Authority: In Defense of Moral Naturalism without Clout', *Philosophical Studies,* (2010), 221-38.

unenlightened desires may frustrate one’s enlightened desires, and vice versa.) Accordingly, Joyce’s conception of the ‘knave’ may be disambiguated as follows:

> **Knave***: a person for whom an immoral action will further his or her *unenlightened* desires, but performing an alternative morally permitted or obligatory action would not.

> **Knave****: a person for whom an immoral action will further his or her *enlightened* desires, but performing an alternative morally permitted or obligatory action would not.

As for the second, Joyce frames his inescapability claim thus:

> **Inescapability**: If *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether he cares to, regardless of whether Φing satisfies any of his desires or furthers his interests.

Again, I must disambiguate this term. Focussing for simplicity just upon desires, when Joyce refers here to *x*’s desires, is he referring to their *enlightened* desires, or to their *unenlightened* (e.g. present) ones? In other words, which of the following claims is Joyce endorsing?

> **Inescapability***: If *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether Φing satisfies any of his *unenlightened* desires.

> **Inescapability****: If *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether Φing satisfies any of his *enlightened* desires.

Since is unclear to which of these Joyce refers (and likewise for the knave), I shall leave both conceptions on the table for now, in case this has any bearing upon the evaluation of the truth of the proposition in premise P1. (There is a third possibility, viz. inescapability***: If *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether Φing satisfies any of his enlightened *or* unenlightened desires. However, this folds into inescapability**, so I shall not address it separately.)

In terms of the third of the above points, it is Goal Theory that I am defending here, so I shall take this to be the theory of morality, as I shall when I evaluate Joyce's authority claim. Remember that, on Goal Theory, *x* morally ought to Φ (in circumstances *C*) just in case doing Φ (in circumstances *C*) would best serve the strongest desire of a fully rational and sufficiently informed version of *x* (i.e. *x*'s true strongest desire). Joyce wants to say that there is probably no version of moral naturalism that either accommodates his practical clout or else denies it without thereby disabling crucial uses of moral concepts.[32] I intend to demonstrate that Goal Theory accomplishes what Joyce thinks cannot be accomplished.

With these things in place, let me now suggest two natural ways of filling out the claim in premise P1, depending upon whether the desires being referenced are unenlightened or enlightened desires:

> **Claim 1**: if morality lacked the feature of inescapability* [i.e. if *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether Φing satisfies any of his *unenlightened* desires], then it would not necessarily be the case that if there is a

[32] Joyce, *The Evolution of Morality*, pp. 192-93.

> moral obligation to do Φ, then a knave* [i.e. a person for whom an immoral action will further his or her *unenlightened* desires, but performing an alternative morally permitted or obligatory action would not] morally ought to do Φ.

> **Claim 2**: if morality lacked the feature of inescapability** [i.e. if *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether Φing satisfies any of his *enlightened* desires], then it would not necessarily be the case that if there is a moral obligation to do Φ, then a knave** [i.e. a person for whom an immoral action will further his or her *enlightened* desires, but performing an alternative morally permitted or obligatory action would not] morally ought to do Φ.

With regard to Claim 1, observe that when we take Goal Theory as our theory of morality, then morality does *not* lack the feature of inescapability*, since, on Goal Theory, what an agent morally ought to do is independent of their unenlightened desires. As such, the antecedent in Claim 1 would be false, and the consequent then immaterial — meaning that Argument 4 would not get started. That said, because Goal Theory accommodates inescapability*, then I would still endorse the corresponding conclusion C2 of Argument 4 (i.e. morality is inescapable*).

So, if we grant that morality is inescapable*, does this get Joyce the categorical reasons conclusion that he wants? No, because if Joyce's conception of inescapability is the one expressed by inescapability*, then (restricting it just to desires, for the purpose of simplicity) premise P1 of Argument 3 becomes:

P1*) if *x* morally ought to Φ, then *x* morally ought to Φ regardless of whether Φing satisfies any of his *unenlightened* desires [inescapability*]

However, notice that this claim, when combined with Joyce's authority claim in premise P2 of Argument 3 (i.e. if *x* morally ought to Φ, then *x* has a reason for Φing), yields the following conclusion:

C1*) therefore, if *x* morally ought to Φ, then *x* has a reason for Φing regardless of whether Φing serves his *unenlightened* desires

However, this is not the categorical reasons conclusion that Joyce requires for his case, because *x*'s reason for Φing might derive from his *enlightened* desires. Thus, any such reason need not necessarily be a *categorical* one.

Now, let me consider Claim 2. Observe in this case that when we take Goal Theory as our theory of morality, then morality *would* lack the feature of inescapability** (and thence inescapability***), because, on Goal Theory, what an agent morally ought to do is defined in terms of their true strongest desire (which is a member of the set of their enlightened desires), and so cannot be independent of their enlightened desires. However, in that case, notice that the knave** is a conceptual impossibility on Goal Theory. By definition, any immoral action for an agent on Goal Theory must frustrate at least one of their enlightened desires (i.e. their true strongest desire), and any moral action for an agent on Goal Theory must further at least one of their enlightened desires (i.e. their true strongest desire), so there can be no agent for whom an immoral action furthers his or her enlightened desires, but performing an alternative morally permitted or obligatory action does not. Thus, Claim 2 (and so premise P1 of Argument 4) becomes incoherent. Moreover, with Goal Theory as our moral theory and inescapability** (or inescapability***) as our conception of inescapability, then conclusion C2 of Argument 4, and thence premise P1 of Argument 3, is false. Thus, once again, Joyce does not get the categorical reasons conclusion that he requires.

In light of that, then depending upon our chosen conceptions of the knave and inescapability, I would argue that Goal Theory either accommodates inescapability, but without this thereby entailing categorical reasons (contra Joyce); or else it denies inescapability, but (contra Joyce) without disabling certain crucial uses of moral concepts, in particular the evaluation of people's actions (since there can then be no knave who can fail to act in accordance with moral evaluations).

Having evaluated Joyce's inescapability claim, let me turn now to his *authority* claim, as expressed in premise P2 of Argument 3 (i.e. if x morally ought to Φ, then x has a reason for Φing). Joyce presents two basic justifications for this claim. Firstly, as mentioned earlier, he invokes Mackie's platitude, viz. it is necessary and *a priori* that, for any agent x, if x ought to Φ, then x has a reason to Φ. Secondly, as part of his case for practical clout being a core desideratum of any moral theory, Joyce suggests that moral concepts could not be used as they characteristically are (in terms of evaluating people's actions), and would thus be abandoned, if knavish individuals had no normative reason to do as they morally ought to do (analogous to his argument for inescapability, as formulated in argument 4).[33]

Now, one might challenge both of the abovementioned arguments. For example, with regard to the first, Shafer-Landau suggests reasons to doubt that Mackie's platitude is a platitude at all.[34] Alternatively, one might grant that this platitude ranges over certain normative domains, but then deny that this applies to the *moral* domain. As for the second, one might argue, for example, that people would continue to make

[33] Joyce, *The Evolution of Morality*, pp. 203-09.

[34] Shafer-Landau, 'Error Theory and the Possibility of Normative Ethics', pp. 110-11.

moral evaluations of others, even in the face of knavish agents who have no normative reason to do as they morally ought to do (a variation on Tresan's response to Joyce's inescapability argument). However, I shall set those objections aside, noting that Goal Theory actually accommodates Joyce's authority claim. On Goal Theory, if *x morally* ought to Φ (in circumstances *C*), then this will be because doing Φ (in circumstances *C*) will best serve the strongest desire of a fully rational and sufficiently informed version of *x* (i.e. *x*'s true strongest desire). Now, the HTR* states that an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. So, on the conjunction of Goal Theory and the HTR* (where the former is committed to the latter, once we grant Proportionalism and Parfit's platitudes), *x* will necessarily have at least *pro tanto* normative reason to do what they morally ought to do. Thus, on Goal Theory and the HTR*, Joyce's formulation of the authority function of morality (and therefore premise P2 of Argument 3) would be affirmed. Note that this conclusion is independent of any desires *x* may *presently* have, since both Goal Theory and the HTR* are formulated in terms of the desires of a fully rational and sufficiently informed version of *x*. Whether and to what extent the things that agents morally ought to do on Goal Theory align with *commonsense* morality (meaning that agents would then have a reason to obey commonsense moral imperatives) is a separate question, and one that I shall return to in section 5.4.

In conclusion, depending upon how one interprets Joyce's inescapability claim, my account is either consistent with both his inescapability and authority claims, yet still does not imply categorical reasons (i.e. I grant premise P2 of Argument 3, as well as one possible reading of premise P1, but then deny that conclusion C1 is thereby entailed); or else it is inconsistent with his inescapability claim (but consistent with his authority claim), yet still does not disable certain crucial uses of moral concepts (contra

Argument 4, and implying that I then deny premise P1 of Argument 3 and thence conclusion C1). In neither case does Joyce get the categorical reasons conclusion that he wants in conclusion C1 of Argument 3. By such means, I would argue that my moral naturalist account plausibly accomplishes what Joyce thinks is not possible, in accommodating a plausible kind of practical clout (i.e. inescapability* + authority, such that if *x* morally ought to Φ, then *x* has a reason for Φing, and *x* morally ought to Φ regardless of whether Φing satisfies any of his unenlightened desires), without disabling certain crucial uses of moral concepts. At the same time, it denies an implausible kind (i.e. inescapability** + authority). Thus, I think that my account plausibly captures the intuition behind Joyce's claim that there is a core desideratum of any moral theory (i.e. inescapable authority), such that a failure to underwrite this is sufficient to warrant rejecting the theory in question — without thereby entailing Joyce's categorical reasons claim. In so doing, it presents a plausible response to Joyce's challenge to moral naturalists, viz. to either accommodate practical clout within a naturalistic framework, or else adduce a cogent denial of it.[35]

Joyce himself acknowledges the rather tentative nature of his defences of the inescapability and authority of morality, saying that philosophers probably have no way to decide whether a commitment to these is a non-negotiable aspect of moral thought and practice, because they just have no settled procedures for determining such matters.[36]

Returning to Joyce's intuition that it would be morally wrong for Jack to murder John, even if murdering John will further Jack's desires, but not doing so will not, we see that Joyce is expressing a moral absolutist view. However, I would submit that my account accommodates these and other similar intuitions. For the kinds of reasons

[35] Joyce, *The Evolution of Morality*, p. 193.

[36] As also noted here: Cuneo, 'Moral Naturalism and Categorical Reasons', p. 114. Cuneo also points out that Joyce arbitrarily assigns weight to certain features of ordinary moral discourse (e.g. holding others morally accountable), whilst discounting others (e.g. stock moral truisms).

already adduced in section 2.5, I think it is very likely in the real world that it would be immoral on Goal Theory for Jack to murder John, insofar as murdering John would be very likely to frustrate Jack's true strongest desire, even if it furthers some of his unenlightened desires (I evaluate challenges to this kind of claim in section 5.4). Moreover, if not murdering John would best serve Jack's true strongest desire (which is one of the desires of a fully rational and sufficiently informed version of Jack), as it would do if this is the moral action for Jack on Goal Theory, then, on the HTR*, Jack would then necessarily have a (decisive) reason to not murder John. As such, we may condemn Jack if he nonetheless murdered John. The only way that we would withhold condemnation of Jack on my account is if, contra what I suggest, it would *not* be immoral on Goal Theory for Jack to murder John (because doing so would best serve his true strongest desire, rather than frustrate it). I explain in section 5.4 why I think this is very unlikely in the real world; but even if I allow it, we would *still* be evaluating Jack's actions in accord with his moral obligations. It is just that his moral obligations in that case would diverge from those of commonsense morality.

Thus, my account yields both authority and what I think is plausible version of inescapability (whilst etiquette, by comparison, only plausibly yields the former), thereby plausibly capturing the intuition behind Joyce's practical clout claim. At the same time, the reason in the authority claim is a *hypothetical*, not a *categorical* one, deriving as it does from its relation to Jack's enlightened desires. Nonetheless, I would suggest that my account might be said to capture the intuition behind Joyce's categoricity claim, in generating a reason that obtains independently of Jack's *unenlightened* desires (we might call this a categorical* reason). At the same time, it does not yield a reason that obtains independently of Jack's *enlightened* desires — where I find such a hypothetical reason to be quite implausible.

As such, I would submit that my account threads the needle. On the one hand, it has sufficient practical clout such that Jack’s murder of John (or Hitler’s order of genocide) would not be regarded as morally right just because Jack (or Hitler) has some unenlightened (e.g. present) desire that would be served by him doing so and no unenlightened desire that would be served by him not doing so (as would appear to be the case on the HTR). On the other hand, it does not yield normative reasons for agents to act in ways that they would only do if they were irrational and/or insufficiently informed (as may be the case if there are categorical normative reasons, obtaining independently of agents’ enlightened desires).

Observe that there are other moral naturalist accounts that do not accomplish what mine does. For example, consider the conjunction of utilitarianism and the HTR. In that case, we might grant Joyce’s inescapability claim. For example, if total utility would be maximised by Jack not murdering John, then the resulting moral obligation applies to Jack independently of his (enlightened or unenlightened) desires. However, that Jack has a moral obligation on utilitarianism to not murder John does not entail that acting thusly would serve any of these desires. Therefore, on the HTR, he need not have a reason to act morally, and so Joyce’s authority criterion is not satisfied. But, in that case, Joyce would argue that moral concepts could not then be used as they characteristically are used (in terms of evaluating people’s actions), and would thus be abandoned, meaning that the account in question would fail to possess what he thinks is a core desideratum of any moral theory.

Finally, let me return to the argument for categorical reasons that I adduced at the beginning of this section, viz.

Argument 2

P1)	If categorical reasons are a non-negotiable element of morality, and if morality is not a fiction, then there are categorical reasons.
P2)	Categorical reasons are a non-negotiable element of morality.
P3)	Morality is not a fiction.
C)	Therefore, there are categorical reasons.

As stated earlier, I endorse premise P3. However, since I argue that Joyce's defence of premise P2 probably fails, then I submit that the conclusion that there are categorical reasons is not shown to be true. As such, I think Goal Theory (with its implied theory of normative reasons, the HTR*) plausibly resists Joyce's argument.

3.3 Shafer-Landau's argument

Shafer-Landau wants to defend the existence of *categorical* reasons, which he understands as reasons that obtain independently of their relation to an agent's commitments (cares, desires, wants, goals, etc.).[37] In so doing, he wants to deny *practical instrumentalism*, which he understands as the view that the only reasons there can be are *hypothetical* reasons, i.e. reasons to do things that are in some way ancillary to the achievement of one's commitments (cares, desires, wants, goals, etc.) Why does Shafer-Landau want to do this? As he says:

[37] Shafer-Landau, 'A Defence of Categorical Reasons', p. 189.

> Apart from the intrinsic interest of the matter, showing that there are categorical practical reasons, and that instrumentalism is false, is important for at least two reasons. First, it would enable us to resist relativistic arguments that assume that moral requirements entail excellent reasons for action, but make reasons contingent on our commitments, thereby making the content of moral requirements contingent on our commitments. Second, it would provide us with an adequate reply to arguments that assume a commitment-independent source of moral requirements, and then proceed, with the help of instrumentalism, to the conclusion that there may be no good reason to abide by morality's demands.[38]

Shafer-Landau rejects Joyce's claim that categorical reasons are a non-negotiable element of morality.[39] Yet he also denies Joyce's claim that there are no such reasons.[40] In order to argue for categorical reasons, Shafer-Landau musters an argument in which he directs our attention to the example of a dedicated, successful immoralist (Shafer-Landau's version of the sensible knave). Shafer-Landau asks us to:

> Imagine a person who is very sharp, very cunning, but also deeply malicious. His happiness is directly proportioned to the misery he wreaks. His top priority in life is to cause pain and suffering, even if, as he knows, such conduct will likely bring an early death, or a long incarceration.[41]

As putative real-world instantiations of this immoralist, Shafer-Landau adduces two examples: (1) an experienced torturer who works on behalf of an authoritarian government, and not only endorses that regime, but also takes active pleasure in breaking his victims; and (2) a person who could easily rescue a young child who has strayed from her parents on a busy street, and is about to toddle into the path of an oncoming car, but instead does nothing and watches in delight as the child in run over and killed. In both cases, and in general, Shafer-Landau suggests that we intuitively

[38] Shafer-Landau, 'A Defence of Categorical Reasons', pp. 189-90.
[39] Shafer-Landau, 'Error Theory and the Possibility of Normative Ethics'.
[40] Ibid.
[41] Shafer-Landau, 'A Defence of Categorical Reasons', p. 190.

regard such an immoralist to be morally obligated to desist from his behaviour. And he asks if we do not also believe that there are excellent reasons for him to refrain, notwithstanding the fact that he has no commitments that would be furthered by him doing so, viz. all the considerations of cruelty and so on that constitute the wrongness of the act.

Thus, Shafer-Landau is attempting here to adduce an example of a possible agent whose commitments are served by perpetrating evil deeds, and who has no commitments that would be served by refraining from so doing. As such, he intends to block the possibility that this agent has any *hypothetical* reasons to refrain. Yet, at the same time, he claims that we would all agree that the immoralist does still have reasons to refrain — thereby entailing that these reasons must be *categorical* ones. Shafer-Landau expresses his argument as follows:

Argument 5

P1)	If there are reasons for these dedicated immoralists to refrain from their evil deeds, then practical instrumentalism is false.
P2)	There are such reasons.
C)	Therefore, practical instrumentalism is false.

This argument turns upon premise P2, so how does Shafer-Landau attempt to justify the proposition therein? Well, he deliberately sets aside a Kantian approach, because he is doubtful that any attempt to show that immoral agents necessarily act irrationally will be successful. As he says:

> Most defenders of categorical reasons, following Kant, have tried to sustain such charges [that instrumentalists exemplify some kind of practical inconsistency in behaviour or commitment]. Their vindication would be welcome news for friends of categorical reasons. But I am not optimistic about this most direct route to instrumentalism's refutation.[42]

As such, Shafer-Landau would concur with Philippa Foot, when she said that we can accuse an evil man of villainy, but not necessarily inconsistency. Although we might frown upon the goal he has set himself, he may be perfectly efficient in achieving it, without thereby sacrificing any of his other goals. To think this is impossible, or that he must be acting contrary to his own commitments if he acts immorally, is unjustified.[43] For reasons that I shall set aside here, I agree with Shafer-Landau that Kantians have failed to show that immoral agents necessarily act irrationally. Having said that, on the plausible assumption that it is irrational to fail to do what a fully rational and sufficiently informed version of oneself will do (which, per statement (S_3) in section 2.1, is what Goal Theory would command), then, on my account, immoral agents *do* necessarily act irrationally. However, since there are only *instrumental* normative reasons on my account (per the HTR*), then this result does not help Shafer-Landau.

Having rejected the Kantian approach, Shafer-Landau adopts an alternative tactic, saying that:

> The cruelties [the immoralist] perpetrates are opposed by a host of considerations that make no mention of his aims. These considerations are reasons—reasons to refrain from deliberately inflicting misery. And these reasons will, first and foremost, mention the suffering of his victims, and the absence of their consent to his treatment. If the immoralist's aversion to being found out enters into it at all, it is only in a subordinate

[42] Shafer-Landau, 'A Defence of Categorical Reasons', p. 205.
[43] Foot, 'Morality as a System of Hypothetical Imperatives'.

> role, as a consideration that may supply an additional reason to refrain from his actions, one that is likelier than the others to motivate him to do the right thing.[44]

He adds that:

> I have tried to reveal its attractions with the examples of the dedicated evildoers. So long as we think—as all of us do—that there are genuine considerations to oppose their cruelty, and also think that such considerations obtain independently of their commitments, then premiss (2) is secure.[45]

If we incorporate these thoughts into Argument 5 — making explicit what is currently implicit in the argument, translating considerations into reasons, targeting the HTR* specifically, and focussing for simplicity upon only one subset of the dedicated immoralists' commitments, viz. their desires — then we may formulate a revised argument as follows:

Argument 6

P1)	If there are genuine reasons for these dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires, then the HTR* is false.
P2)	If we all think that there are genuine reasons for these dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires, then there are genuine reasons for these dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires.

[44] Shafer-Landau, 'A Defence of Categorical Reasons', pp. 191-92.
[45] Shafer-Landau, 'A Defence of Categorical Reasons', p. 192.

P3)	We all think that there are genuine reasons for these dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires.
C1)	Therefore, there are genuine reasons for these dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires.
C2)	Therefore, the HTR* is false.

In order to assess the soundness of this argument, a disambiguation of what Shafer-Landau means by the dedicated immoralists' *desires* is called for. Here is one plausible way to do this, based upon whether the desires being referenced are the immoralists' unenlightened or enlightened ones:

> **Desires***: the dedicated immoralists' *unenlightened* (e.g. present) desires.
>
> **Desires****: the dedicated immoralists' *enlightened* desires.[46]

If Shafer-Landau means desires*, then I would deny premise P1, since the antecedent does not entail the consequent. If there are reasons for dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires*, then these reasons may derive from serving their *enlightened* desires — something that is entirely compatible with the HTR*. Thus, with premise P1 being false, C2 is not shown to be true, and the HTR* is left standing.

[46] A third possibility should be mentioned, viz. desires***: the dedicated immoralists' unenlightened *and* enlightened desires. However, this case folds into that for desires**.

If Shafer-Landau instead means desires**, then I would now accept premise P1, but would deny premise P2 (something I would do in the previous case too) and thence conclusion C1. Observe that premise P2 is of the following basic form:

> **(A)**: if we all think that *p* [e.g. there are genuine reasons for the dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires**], then *p*.

However, (A), as presented by Shafer-Landau, is an undefended assertion. I might leave it there, with premise P2, and thence conclusions C1 and C2 being unproven. However, adopting the principle of charity, I shall consider whether (A) might be justified somehow. Perhaps there is established empirical evidence supporting the general reliability of people's beliefs on the truth of such propositions, for example. I am not optimistic, however. Rather, I would suggest that any justification for (A) would likely depend upon our *intuition* — in particular our *moral* intuition. Shafer-Landau actually advocates a moral epistemology incorporating what he describes as a form of Rossian intuitionism, suggesting that he would probably be content to rely upon moral intuition here.[47] In that case, we may rewrite (A) as follows:

> **(A*)**: if we all have an *intuition* that *p* [e.g. there are genuine reasons for the dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their desires**], then *p*.

As such, the source of the 'evidence' that there are reasons, obtaining independently of their commitments, for the dedicated immoralist to refrain from his or her evil deeds

[47] Shafer-Landau, *Moral Realism: A Defense*, pp. 229-302.

(call it e_I) would ultimately be our moral intuition. However, I would argue that (A*) is epistemically problematic, and ultimately unjustified, insofar as our moral intuition is unreliable, and we have no generally accepted means to distinguish any trustworthy moral intuitions from untrustworthy ones. Shafer-Landau would presumably resist that claim, so how would I defend it? Let me explain, by undertaking an excursus on the subject of moral intuitions. The results of such an excursus will also prove useful in subsequent chapters (especially chapter 5), so I shall go into some depth.

3.4 The unreliability of moral intuitions

Many philosophers rely extensively upon intuitions when doing philosophy, with such intuitions often seen as being analogous to observations in science (filling the roles of data and confirmers or falsifiers of theories).[48] However, there are plausible reasons to think that our intuitions in general are unreliable (or, at least, not reliable *enough* to act as justifiers), and therefore that any beliefs based solely upon such intuitions are not justified.[49]

We know, for example, that intuitions sometimes contradict one another, are contradicted by empirical evidence, and vary between people and groups of people (and can sometimes conflict within the same person).[50] Even when there is widespread and stable agreement upon the content of intuitions, these intuitions have routinely turned

[48] As Talbot points out: Brian Talbot, 'Psychology and the Use of Intuitions in Philosophy', *Studia Philosophica Estonica,* 2 (2009), 157-76 (p. 157).

[49] On that point, see, for example: Talbot, 'Psychology and the Use of Intuitions in Philosophy'. In addition, Robert Cummins argues that concerns about their accuracy (amongst other things) render such intuitions 'epistemologically useless': Robert Cummins, 'Reflections on Reflective Equilibrium', in *Rethinking Intuition,* ed. by M. DePaul and W. Ramsey (Lanham, Md: Rowman & Littlefield, 1998), pp. 113-28. And Hilary Kornblith says that intuitions merely tell us about our concepts, and that 'philosophy cannot live up to its ambitions' if it continues to emphasize the use of intuitions: H. Kornblith, 'Appeals to Intuition and the Ambitions of Epistemology', in *Epistemology Futures,* ed. by S. Hetherington (Oxford: Oxford University Press, 2006), pp. 10-25.

[50] On the last point, see for example: J. Weinberg, S. Nichols, and S. Stich, 'Normativity and Epistemic Intuitions', *Philosophical Topics,* (2001), 429-60.

out to be wrong, e.g. the ubiquitous but faulty intuitions that the Earth is flat and does not move, that the sun moves around the Earth, that slavery is morally acceptable, and that homosexuality is morally wrong (the last two are still endorsed in some quarters, but generally regarded by moral philosophers as being false). Even today, our intuitive theories about the world are often wrong (on topics as diverse as mechanics, essentialism, thermodynamics, germ theory, evolution, matter, teleology, animism, and astronomy), and persist even when we have explicitly rejected them.[51] Moreover, no reason is generally given why we should accept specific intuitions as evidence; and we have no generally accepted means to distinguish trustworthy intuitions from untrustworthy ones.[52]

With regard to moral intuitions in particular (which, in line with Walter Sinnott-Armstrong, I shall understand as being strong and immediate moral beliefs[53]), we often find ourselves unreflectively inclined to accept them as true. As Shelly Kagan observes:

> Given a conflict between a theory — even one that seems otherwise attractive — and an intuitive judgment about a particular case that conflicts with that theory, we will almost always give priority to the intuition.[54]

However, I shall argue that moral intuitions are vulnerable to a number of serious objections, collectively rendering them *unreliable*.

The sense of reliability on which moral intuitions are being attacked is not the *baseline accuracy* sense, on which moral intuitions would be reliable if, on balance,

[51] See, for example: Andrew Shtulman and Kelsey Harrington, 'Tensions between Science and Intuition across the Lifespan', *Topics in Cognitive Science,* (2016), 118–37; Andrew Shtulman and Joshua Valcarcel, 'Scientific Knowledge Suppresses but Does Not Supplant Earlier Intuitions', *Cognition,* (2012), 209-15.

[52] Talbot, 'Psychology and the Use of Intuitions in Philosophy', p. 158.

[53] Walter Sinnott-Armstrong, 'Framing Moral Intuitions', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity,* ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2007).

[54] Shelley Kagan, 'Thinking About Cases', in *Moral Knowledge,* ed. by Ellen Frankel Paul, Jr. Miller, Fred, and Jeffrey Paul (Cambridge: Cambridge University Press, 2001), pp. 44-63 (p. 44).

they are right more often than wrong (Boyd and Nagel, for example, defend the reliability of intuitions on this sense). After all, it seems implausible to claim that some belief-generating mechanism would be a reliable one if it were to marginally improve upon a coin toss. Instead, the most relevant sense of reliability here is one of *trustworthiness,* which requires a correspondingly higher threshold for the propensity of true deliverances, such that a moral intuition being *reliable* in this sense would be sufficient for a belief based on that intuition to be epistemically justified. Accordingly, I shall adopt that sense here. I leave the precise threshold unspecified, but take it to be significantly higher than 0.5 (though less than 1, since reliability does not require infallibility).[55] If moral intuitions fail to meet this threshold, then they should probably be excluded from being considered reliable evidence directly relevant to what our theories of normative reasons need to explain.

It may be objected that if we judge moral intuitions unreliable, then it is difficult to imagine how to develop and assess a moral theory. In reply, I would point to the approach and dialectic that I adopt. This has some similarities to that adopted by Peter Singer — described by him as taking a top-down approach to moral theorising, selecting ‘a theory that is based on a fundamental axiom that seems ... clear and undeniable’, and then applying the theory to particular situations, accepting whatever conclusions it generates. In that case, permitting moral intuitions to serve as a test of a moral theory would be to deprive the theory of its critical capacity.[56] In my case, the equivalent of Singer’s fundamental axiom would be the proposition (S_1) in section 2.1, viz. if there is a true moral system, then its system of imperatives supersedes all other imperatives for rational agents. I do not claim that this proposition is self-evidently

[55] This accords with experimentalist critiques of intuitions e.g. Joshua Alexander and Jonathan M. Weinberg, 'The “Unreliability” of Epistemic Intuitions', in *Current Controversies in Experimental Philosophy,* ed. by E. Machery and E. O’Neill (Oxford: Routledge, 2014), pp. 128-45. Perhaps my arguments motivate the conclusion that moral intuitions are unreliable on the baseline accuracy sense of reliability too, but my case does not rest upon justifying that stronger claim.

[56] Peter Singer, 'Sidgwick and Reflective Equilibrium', *Monist,* (1974), 490-517 (p. 516).

true. Rather, I suggest it is underwritten by an appeal to the community’s linguistic intentions, and so is an *a posteriori* observation of how people use moral language. As such, I would argue that it does not rely upon moral intuition in any substantive way. The same applies to the subsequent steps in the derivation of Goal Theory.

Specifically, the move from proposition (S_1) to (S_2) is based upon a semantic equivalence; that from (S_2) to (S_3) is based upon a *reductio ad absurdum*; that from (S_3) to (S_4) upon an appeal to the widely accepted action-based theory of desire; and that from (S_4) to (S_5) is once again based upon a semantic equivalence. At no point is there any fundamental dependence upon a strong and immediate moral belief (i.e. a moral intuition). In terms of assessment, I am evaluating Goal Theory in terms of its theoretical adequacy and resistance to some dominant objections. Once again, I think it is implausible to claim that this evaluation is fundamentally dependent upon strong and immediate moral beliefs. Like Singer, I then apply the theory to particular situations, accepting whatever conclusions it generates, having more justified trust in the correctness of the theory than in any conflicting moral intuition.

My goal here is not to construct the definitive case against moral intuitions, as that would consume at least a thesis in itself. Instead, I have the more modest aim of casting sufficient doubt upon the reliability of moral intuitions that we may have justifiable scepticism regarding (A*), and thus reasonably deny that moral intuitions constitute reliable evidence with which our candidate theories of normative reasons should fit. My primary argument may be formalised as:

Argument 7

P1)	Probably, one is justified in believing on the (sole) basis of a putative source of evidence only if one lacks (undefeated) reason to think it unreliable.
P2)	We probably have (undefeated) reasons to think moral intuitions unreliable.
C1)	Therefore, beliefs based (solely) on moral intuitions are probably not justified.
P3)	If beliefs based (solely) on moral intuitions are probably not justified, then moral intuitions should probably be excluded from being considered reliable evidence directly relevant to what our theories of normative reasons need to explain.
C2)	Therefore, moral intuitions should probably be excluded from being considered reliable evidence directly relevant to what our theories of normative reasons need to explain.

I assume the premise most likely to be challenged is P2, but before I defend that premise I should first address the possible objection that premise P1 itself depends upon an appeal to epistemic intuitions, thereby rendering the argument self-undermining. In response, I would point out that if we have (undefeated) reason to think that a source of evidence is *unreliable* (in the trustworthiness sense that I am employing here), then we have (undefeated) reason to think it will fail to accord with reality sufficiently often for a belief based on that evidence to be epistemically justified. In that case, one would not be justified in believing on the (sole) basis of that source of evidence. One would only be justified in believing on the (sole) basis of that source of evidence if we lacked (undefeated) reason to think that the source of evidence was unreliable — as premise P1 asserts.

With regard to my defence of premise P2, I shall argue that we probably have (undefeated) reasons to think moral intuitions are unreliable. In particular:

1. Moral intuitions suffer from pervasive and persistent interpersonal intrasource inconsistency or disagreement — including amongst thoughtful, reflective, and comprehending people, such as moral philosophers.
2. Many moral intuitions are vulnerable to framing effects, biases, and suchlike.
3. To the extent that moral intuitions are tracking something, it does not appear to be an independent moral reality, but instead the kind of behaviours that tended to increase the differential reproductive success of our Pleistocene hunter-gatherer ancestors (where the two are not plausibly extensionally equivalent).
4. These problems can be at best only partly ameliorated by means of the application of a process of reflective equilibrium.
5. There are good reasons to be sceptical of the intuitionist claim that certain moral propositions are *self-evident*, thereby necessitating no additional proof.

I shall now discuss each of these points in turn.

First, there are many cases where one has the moral intuition that *p*, yet another person either fails to have the intuition that *p*, or else has an intuition whose content is ~*p* (or has an intuition that is not the explicit propositional negation of *p*, but which can be shown to contradict *p* once we incorporate some other justified principle).[57] In that

[57] This *interpersonal intrasource inconsistency* contrasts with *intrapersonal intersource* inconsistency, whereby one has an intuition-independent justification for thinking that ~*p*, whilst one's intuition is *p*; with *intrapersonal intrasource inconsistency,* where one sometimes finds *p* intuitive and sometimes finds

case, many moral intuitions must be false (by the principle of non-contradiction); and the correct response in light of such disagreement about *p* will typically be a suspension of belief or an appropriate reduction in credence.[58] In that case, we already have reason to consider moral intuitions unreliable. Some of these disagreements may be eliminated on the basis that they are not epistemically significant, insofar as one or both of the parties concerned has failed to properly grasp the relevant propositional content, is insincere, or is otherwise lacking in some necessary competence. However, even setting such cases aside, there is still a great deal of interpersonal intrasource inconsistency. For example, if we restrict our attention to moral philosophers, then we should hope to have a set of individuals who have a clear understanding of the moral propositions in question, have thought long and hard about them, and are sincere. Yet one has only to consult the pages of an anthology of applied ethics in order to bear witness to pervasive disagreement regarding the moral intuitions of moral philosophers. Of course, some of this disagreement will arise from them endorsing different normative theories. However, one could argue that even one's disposition to prefer one or other normative theory is to some significant extent influenced by one's moral intuitions.

The exact nature and extent of the aforementioned interpersonal intrasource inconsistency is a matter for empirical research — and such research is ongoing. For example, the extent of variation in intuitions between different persons, groups, and so on, is the subject of the *variation project* in experimental philosophy. Some preliminary results are already being cited as supporting an argument from interpersonal intrasource

~*p* intuitive; and with *interpersonal intersource* inconsistency, on which one has the intuition that *p* but some other person has intuition-independent justification for believing that ~*p*. These others may also constitute a challenge for moral intuitions, but I think that the greatest challenge is likely to come from *interpersonal intrasource* inconsistency, which is why I focus upon it here.

[58] E.g. L. BonJour, *In Defense of Pure Reason* (Cambridge: Cambridge University Press, 1998), pp. 138-42.

inconsistency.[59] I am willing to sign on to the claim that there is some (relatively small) subset of moral intuitions about which (almost) everyone would agree (the ‘stock moral truisms’ referred to in section 2.5, e.g. ‘the deliberate humiliation, rape, and torture of a child, for no purpose other than the pleasure of the one inflicting such treatment, is immoral’[60]). However, this still leaves all of the other moral intuitions vulnerable to the problem of interpersonal intrasource inconsistency or disagreement.

As a result, we appear to have some (relatively large) subset of moral intuitions about which we find interpersonal intrasource inconsistency or disagreement amongst informed and competent interlocutors, and thus a suspension of belief or reduction in credence is an appropriate response — thereby plausibly rendering them unreliable. And we have some other (relatively small) subset of moral intuitions about whose propositional content (almost) all informed and competent interlocutors agree. Yet the absence of disagreement does not thereby entail the *reliability* of these moral intuitions. A moral intuition might garner almost universal agreement (including amongst philosophers), yet still be *wrong* (such as once prevalent intuition that slavery is acceptable). Whether the moral intuitions in this second subset should be justifiably considered reliable will depend upon whether we have some independent justification for believing that they are. Here I am not making a demand that is in principle unsatisfiable, viz. that intuition must be calibrated by some other source, which must in turn be calibrated by yet another source of evidence, which must then be calibrated by

[59] See, for example: Joshua Alexander, Ronald Mallon, and Jonathan M. Weinberg, 'Accentuate the Negative', *Review of Philosophy and Psychology,* 2 (2010), 297–314; Wesley Buckwalter and Stephen Stich, 'Gender and Philosophical Intuition', in *Experimental Philosophy,* ed. by Joshua Knobe and Shaun Nichols (Oxford: Oxford University Press, 2014); Stacey Swain, Joshua Alexander, and Jonathan M. Weinberg, 'The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp', *Philosophy and Phenomenological Research,* 76 (2008), 138-55; Weinberg, Nichols, and Stich, 'Normativity and Epistemic Intuitions'. Concerns have been raised about survey methods, but even those critical of them concede that ‘there will definitely be a *prima facie* problem for the appeal to intuitions in philosophy if surveys show that there is extensive enough disagreement on the subject matter supposedly open to intuitive access.’ Ernest Sosa, 'Experimental Philosophy and Philosophical Intuition', *Philosophical Studies,* 132 (2007), 99-107.

[60] R. Shafer-Landau, 'Defending Ethical Intuitionism', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity,* ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2008), pp. 83-96 (p. 83).

another, and so on ad infinitum. Rather, I regard some sources of evidence as being properly basic (and thus self-justifying), and think that these can serve as the basis for our non-foundational beliefs. (I would deny that moral intuitions themselves can be similarly self-justifying, but will set that question aside for the moment.)

So, it seems that the set of moral intuitions from competent and sincere interlocutors, *S*, can plausibly be divided into two subsets: (1) a (larger) subset, S_1, of those that are probably unreliable; and (2) a (smaller) subset, S_2, of moral intuitions about whose reliability we should remain agnostic (in the absence of any independent justification for believing that they are either reliable or unreliable). This already casts doubt upon the claim that moral intuitions constitute reliable evidence with which our candidate theories of normative reasons should fit, but there are more problems to come.

Second, research suggests that very many moral intuitions are systematically distorted by philosophically irrelevant factors, such as framing effects and biases.[61] For example, as Alexander & Weinberg found, people's intuitions about such moral dilemmas as the trolley problem differ depending upon whether they are told them before or after other dilemmas, or in the first or third person, for example.[62] One study also showed that the willingness of people to agree that it was morally permissible to sacrifice an innocent man in order to save a greater number of people varied according to the conjunction of the respondents' political alignment and the perceived race of the

[61] E.g. Walter Sinnott-Armstrong, 'Moral Intuitionism Meets Empirical Psychology', in *Metaethics after Moore*, ed. by Terry Horgan and Mark Timmons (Oxford: Oxford University Press, 2006), pp. 339-66 (p. 353).

[62] Joshua Alexander and Jonathan M. Weinberg, 'Analytic Epistemology and Experimental Philosophy', *Philosophy Compass*, 2 (2007), 56–80. For more on this problem, see: Swain, Alexander, and Weinberg, 'The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp'. Also: J. Haidt and J. Baron, 'Social Roles and the Moral Judgement of Acts and Omissions', *European Journal of Social Psychology*, 26 (1996), 201– 18; A. Tversky and D. Kahneman, 'The Framing of Decisions and the Psychology of Choice', *Science*, 211 (1981), 453– 58.

innocent man.[63] These kinds of worries, in and of themselves, might be sufficient to render moral intuitions generally unreliable, on the basis that we are justified in believing that a large percentage of moral intuitions are false. Sinnott-Armstrong, for example, offers an argument to this effect.[64] He grants that some moral intuitions can justify moral beliefs. However, in light of the fact that the justification offered by so many of them is undermined by distorting factors, it would need to be determined that any particular moral intuition is not one of the undermined ones before we can accept that it provides justification. (He goes on to argue that any intuitions that do provide justification would do so only *inferentially*, and so there is no *non*-inferential justification for our moral beliefs — and hence moral intuitionism is false.)

Even when the moral intuitions concerned are ones about which (almost) everyone would agree, that a large percentage of moral intuitions are probably false does still pose a challenge for the proponent of moral intuitions. As Sinnott-Armstrong says:

> The evidence from framing effects does cast initial doubt on the reliability of such uncontroversial moral beliefs insofar as the evidence creates a presumption that needs to be rebutted. Because such beliefs fall into a class with a large percentage of falsehoods, it is reasonable to ascribe that same probability of falsehood unless and until the believer has reason to believe that such uncontroversial moral beliefs are more reliable than average.[65]

Perhaps the presumption can be successfully rebutted in some cases, but the onus is upon the advocate of moral intuitions to do so.

[63] See: David A. Pizarro and others, 'The Motivated Use of Moral Principles', *Judgment and Decision Making*, 4 (2009), 476-91.

[64] Sinnott-Armstrong, 'Framing Moral Intuitions'; Sinnott-Armstrong, 'How to Apply Generalities: Reply to Tolhurst and Shafer-Landau'.

[65] Sinnott-Armstrong, 'How to Apply Generalities: Reply to Tolhurst and Shafer-Landau', p. 105.

In addition to the above-mentioned framing effects, Josh Greene has also argued that unreliable mental processes cause some of our moral intuitions, giving us good reason to discount these moral intuitions.[66] On Dual Process Theory, we have two different cognitive systems in our brains, viz. System 1 and System 2 — with the former being emotional, automatic, and quick; and the latter being controlled, deliberate, and slow (what we normally consider to be conscious reasoning). Greene argues that the presence of these two systems explains why altering morally irrelevant factors in a thought experiment may generate conflicting moral intuitions.

For example, he theorises that when people are asked to imagine pushing a large man from a footbridge onto a railway track to prevent a runaway trolley from killing five innocent workers, System 1 will generate the intuition (disinclining people from pushing the man). By contrast, when people are asked to imagine pulling a switch to open a trapdoor through which the same large man will fall onto the track and stop the trolley, System 2 generates the intuition (this time inclining people to pull the switch). Greene suggests that the two systems generate conflicting intuitions because System 1 produces characteristically *deontological* judgements, whilst System 2 produces characteristically *consequentialist* judgements (where the two yield conflicting imperatives in certain situations). Greene then argues that when we make some plausible normative assumptions about what constitutes a good judgement (e.g. not being biased by irrelevant emotional information), we should not trust those moral intuitions produced by System 1.

In light of the foregoing, even further doubt is cast upon the already likely unreliable subset of moral intuitions, S_1, with the problem of interpersonal intrasource inconsistency or disagreement being compounded by that of vulnerability to framing effects and biases. Moreover, the evidence of framing effects also casts doubt upon the

[66] J. Greene and others, 'Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment', *Cognition*, 111 (2009), 364– 71.

reliability of the moral intuitions in S_2, with the onus being placed upon their defender to adduce reasons to believe that these intuitions are more reliable than the average.

Thirdly, I would argue that moral intuitions do not reliably track moral truth. Yet, if moral intuitions do not track the moral truths that they claim to track, then moral intuitions in general will be unreliable sources of moral knowledge, even if everyone agrees about their propositional content, and the evidence of framing effects and biases does not defeat them. In that case, even if the propositional content of some moral intuitions is true, it will be so only accidentally, with it not being appropriately connected to moral truth (so even if some such proposition *p* is true — e.g. torturing children for fun is morally wrong — it is not *because* of the fact that *p* is true that the content of our moral intuition is *p*; and were it not the case that *p* is true then the content of our moral intuition would not be *p*).

Why think that moral intuitions do not reliably track moral truth? Well, I would argue that our innate moral sense, from which our moral intuitions arise, is very likely an evolutionary adaptation (or by-product of this) that would have tended to increase our Pleistocene hunter-gatherer ancestors' differential reproductive success — in part by motivating and reinforcing (through such reactive emotions as guilt and resentment) a set of prosocial behaviours from which individuals reaped the benefits of direct and indirect reciprocity, avoided sanctions, and reinforced the prosocial behaviour of others with whom they interacted.[67] This psychological altruistic disposition would then have

[67] On the evolutionary basis of our innate moral sense, see for example: L. Cosmides and J. Tooby, 'Cognitive Adaptations for Social Exchange', in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture,* ed. by J.H. Barkow, L. Cosmides, and J. Tooby (Oxford: Oxford University Press, 1992); Frans De Waal, *Primates and Philosophers: How Morality Evolved*, ed. by Stephen Macedo and Josiah Ober (Princeton, NJ: Princeton University Press, 2009); J. Haidt and C Joseph, 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues', *Daedalus: On Human Nature,* 133 (2004), 55-66; M. D. Hauser, *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong* (New York: Ecco Press, 2006); Joyce, *The Evolution of Morality*; Philip Kitcher, 'Biology and Ethics', in *The Oxford Handbook of Ethical Theory,* ed. by D. Copp (Oxford: Oxford University Press, 2005); M. Ridley, *The Origins of Virtue: Human Instincts and the Evolution of Cooperation* (London: Penguin, 1997); E. Sober and D. S. Wilson, *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998); Robert L. Trivers, 'The Evolution of Reciprocal Altruism', *The Quarterly Review of Biology,* 46 (1971), 35-57.

been attenuated by factors from thousands of years of cultural evolution (such as those ‘from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for the survival of the group in social and economic circumstances that now lie in the distant past’).[68] We would then be left with a system of moral intuitions ‘thoroughly saturated with evolutionary influence’, and further distorted by cultural factors.[69]

I am inclined to agree with Walter Sinnott-Armstrong, Liane Young, and Fiery Cushman that moral intuitions may plausibly be considered to be a motivational heuristic for producing certain kinds of prosocial evolutionarily adaptive behaviour — where this is a particular instance of a general *affect* heuristic, on which if thinking about an act makes you feel bad in a certain way, then you judge that it is morally wrong (specifically, moral intuitions are ‘subjective psychological states that exist because they motivate fitness-enhancing behaviors in a computationally efficient manner’).[70] However, I then disagree with Sinnott-Armstrong et al. that *moral truth* is the target attribute of the heuristic (where the target attribute is held to be relatively inaccessible, the heuristic attribute [i.e. moral intuitions] are much more easily accessible, and there is plausibly an unconscious substitution of the target attribute for the heuristic attribute). Instead, and in line with the above, I would suggest that the target attribute is more probably what we might call ‘smoral’ truth, where this is a set of truths deriving from the kind of prosocial behaviours that would have tended to increase the differential reproductive success of our Pleistocene hunter-gatherer

[68] As proposed in: Philip Kitcher, 'Between Fragile Altruism and Morality: Evolution and the Emergence of Normative Guidance', in *Evolutionary Ethics and Contemporary Biology,* ed. by G. Boniolo and G. De Anna (Cambridge: Cambridge University Press, 2006), pp. 159–77. Quote from: Singer, 'Sidgwick and Reflective Equilibrium', p. 516.

[69] S. Street, 'A Darwinian Dilemma for Realist Theories of Value', *Philosophical Studies,* 127 (2006), 109–66 (p. 114). Street (as well as Joyce) thinks that the evolutionary accounts of the origin of our moral sense undermine moral realism (see: Joyce, *The Evolution of Morality*; Street, 'A Darwinian Dilemma for Realist Theories of Value'.) I reject such evolutionary debunking arguments, but will set that aside here.

[70] See: Walter Sinnott-Armstrong, Liane Young, and Fiery Cushman, 'Moral Intuitions', in *The Moral Psychology Handbook,* ed. by John M. Doris (Oxford: Oxford University Press, 2010), pp. 246-72 (p. 262).

ancestors. This need not align with moral truth, unless we implausibly define moral truth in terms of differential reproductive success.

In light of the foregoing, there are at least two significant problems with the claim that moral intuitions reliably track moral truth. Firstly, even if the subset of prosocial behaviours that would have tended to increase our Pleistocene hunter-gatherer ancestors' differential reproductive success perfectly aligned with what is morally good or right, our moral intuition is only a crude and indirect means to motivate that evolutionarily adaptive behaviour, operating non-rationally and from incomplete or false information. Thus, even on the aforementioned assumption, many moral intuitions may still fail to align with moral truth.

Secondly, why imagine that the subset of prosocial behaviours that would have tended to increase our Pleistocene hunter-gatherer ancestors' differential reproductive success is perfectly aligned with what is morally good or right anyway? It is plausible that the two are aligned to some extent in certain circumstances (with regard to our stock moral truisms, for example). After all, it seems that dispositions towards cooperation and psychological altruism would have helped in some circumstances to improve the differential reproductive success of our ancestors (and vice versa). This is further supported by the results of game theory, on which cooperative and reciprocally altruistic interactions may lead to *positive-sum* games, in which all of the interested parties gain.[71] And our plausible accounts of e.g. moral rightness (e.g. contemporary consequentialist, Kantian, virtue, and contract theories of morality, along with Goal Theory), with their corresponding moral concepts of fairness, justice, impartiality, guilt, and so on, would tend to endorse acting accordingly.

At the same time, with the transition to larger groups, we also evolved strong dispositions towards loyalty to our ingroup, along with a corresponding dislike,

[71] See, for example: Binmore, *Natural Justice*. Also: Drescher, *Good and Real: Demystifying Paradoxes from Physics to Ethics*, pp. 273-320.

suspicion, or demonization of outgroups; obedience and deference to tradition and authority; conformity to social and community norms; and an abhorrence for certain things perceived to be impure or disgusting, perhaps including behaviours such as homosexuality (with different people and cultures according differing weights to these elements, as well as to those of care, fairness, and liberty — which explains some of the pervasive interpersonal intrasource inconsistency or disagreement that we find with moral intuitions).[72] These dispositions may have had some benefits, including helping to reinforce the altruistic tendencies promoting social cohesion and stability, but they have also undoubtedly motivated much behaviour that has caused great harm in the world, including persecution, oppression, murder, war, and genocide (and our plausible accounts of moral rightness would tend to *condemn* these sorts of actions). [73] Thus, I would argue that the behaviours that would have tended to increase our Pleistocene hunter-gatherer ancestors' differential reproductive success are often rather poorly aligned with what is morally good or right (albeit aligning much better in some paradigmatic cases).

Accordingly, with our moral intuition only approximately tracking the subset of prosocial behaviours that would have tended to increase our Pleistocene hunter-gatherer ancestors' differential reproductive success, and this evolutionarily adaptive behaviour only sometimes aligning with what is morally good or right, I would argue that in general our moral intuition does not reliably track moral truth. As such, the case for the unreliability of moral intuitions (which is already looking robust based upon the

[72] See, for example: Jonathan Haidt, 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', *Psychological Review*, 108 (2001), 814-34. Also: Kitcher, 'Between Fragile Altruism and Morality: Evolution and the Emergence of Normative Guidance'.

[73] For examples of the kind of harms produced, and the corresponding reductions in these harms as we have more recently worked *against* some aspects of our intuitive moral sense, by avoiding undue moralisation, especially where this relates to authority, conformity, and purity, see: Steven Pinker, *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes* (London: Penguin Books, 2011).

problem of interpersonal intrasource inconsistency or disagreement and the evidence of framing effects and biases) is strengthened still further.

Sharon Street also presents an evolutionary argument against the reliability of our intuitive moral judgements, whereby such judgements reflect our evolved natures and dispositions, with no reason to think they would be sensitive to moral truths, making it a massive coincidence if these aligned.[74] This argument poses a particular epistemological challenge to non-naturalist views, such as Shafer-Landau's (one of her main targets), because these views typically rely upon a faculty of intuition to gain reliable moral knowledge.[75] Yet if this faculty of intuition is unreliable, and any correct judgements are only correct by accident, then it seems hardly capable of giving us moral knowledge.

Notwithstanding the foregoing, many who endorse moral intuitions would say that our intuitions are still getting at something important that needs to be taken into account, and that the intuitions we should rely upon are not immediate gut reactions to things, but instead 'considered judgements' (i.e. the judgements we make about situations after some reflection on what is relevant and what is not relevant). According to intuitionists, moral intuitions must be inputs in our search for moral knowledge, because they are our access to the independent moral reality. These ideas are central parts of the predominant methodology in moral theory, viz. wide reflective equilibrium. On wide reflective equilibrium, we evaluate normative theories by bringing into equilibrium ordinary judgements or intuitions about particular cases, moral principles, and background (philosophical and scientific) theories.[76] By means of a successful application of the process, we aim at a position whereby our considered intuitions are

[74] Street, 'A Darwinian Dilemma for Realist Theories of Value'.

[75] Street's argument also poses a challenge to moral realism in general (in terms of either denying any relationship between how we have evolved and what the moral facts are, or else explaining how these things are related), but I shall set that aside here.

[76] For more on this method, see: N. Daniels, 'Wide Reflective Equilibrium and Theory Acceptance in Ethics', *The Journal of Philosophy*, 76 (1979), 256– 82; J. Rawls, 'Outline of a Decision Procedure for Ethics', *The Philosophical Review*, 60 (1951), 177– 97.

fully in harmony (in terms of internal coherence) with our considered moral principles. Reflective equilibrium gives intuitions automatic credibility as inputs to our moral deliberations, but we may not be able to save all of our intuitive judgements, and some of our principles may need to be modified or thrown out altogether. Would an application of this process resolve the problems facing moral intuitions that I have identified? I would submit not.

I have argued that pervasive interpersonal intrasource inconsistency or disagreement, compounded by the evidence of framing effects and biases, and further exacerbated by a failure to reliably track moral truth, probably renders moral intuitions unreliable. However, the proponent of wide reflective equilibrium can respond that considered judgements are the moral intuitions that we have after correctly reflecting upon, and thence eliminating, any inconsistency, framing effects, and so on.[77] Therefore, these intuitions are not unreliable, and so it is then these intuitions to which we should attend. However, there are a number of challenges facing such a view.

Firstly, whilst correct reflection may enable us to partly ameliorate the problems caused by disagreement and framing effects, I am very sceptical that we can resolve them all. As explained earlier, I think there will be many intuitions about whose propositional content we will find significant interpersonal intrasource inconsistency or disagreement amongst (equally) informed, competent, and sincere interlocutors (thereby rendering such intuitions *prima facie* unreliable). In that case, unless we have independent support for some of these intuitions, or we know that some of them, but not others with which they are inconsistent, were formed in a way likely to make them defective, then, with no generally accepted way to distinguish any trustworthy intuitions from untrustworthy ones, the natural conclusion of a process of reflective

[77] Despite requiring us to give thought to the cases at hand, these are still classed as intuitions, since they are arrived at non-inferentially. By ‘correctly’ reflecting, Rawls meant doing so calmly and with adequate information.

equilibrium may be to eliminate them all, on the basis that some significant proportion will be false, but we do not know which ones.

Moreover, in the absence of independent support, those about which there is no disagreement are still not known to be reliable. If there were no evidence for or against their truth, then they would have a 50% probability of being false. However, with no generally accepted way to distinguish the trustworthy ones from the untrustworthy, they all become epistemically dubious. So, should we eliminate all of these too (leaving us with *no* considered judgements), or keep them all (despite them being epistemically dubious)? I think the answer is unclear. And with regard to framing effects, as Sinnott-Armstrong observes, the evidence suggests that reflection can remove some, but not all of these (and the same may be true for Greene's evidence of the cognitively defective way in which some intuitions are formed).[78]

Secondly, it is unclear to me how we can use a process of reflective equilibrium to eliminate the unreliability that accrues from moral intuitions' evolutionary and cultural provenance. I have argued that moral intuitions are effectively pre-theoretical views that probably result from the conjunction of our evolutionary heritage and distorting cultural factors (including bias, superstition, and historical accident). And since we do not understand moral truth simply as a function of what helped Pleistocene hunter-gatherers to maximise their reproductive output (and moral intuitions plausibly do not even reliably track that), then they are arguably all produced in a way likely to make them defective. As such, they probably have no evidentiary plausibility, and should therefore play no role in the construction or justification of our moral theories (and I have endeavoured to respect that thought in the construction and justification of Goal Theory). This would distinguish moral intuitions from observations in science, for example, which are generally (if not universally — there are, after all, scientific anti-

[78] Sinnott-Armstrong, 'Framing Moral Intuitions', p. 70.

realists) taken to reliably track *physical* reality, even if no evidence is entirely free of theoretical contamination. However, simply making coherent a set of beliefs with no initial credibility does not yield justification, because coherent falsehoods are falsehoods nonetheless.[79] Thus, I would argue that moral intuitions or judgements should be afforded no initial credibility in any process of reflection.

The proponent of reflective equilibrium could argue that we may identify the reliable intuitions on the basis that these are the ones in harmony with our plausible moral theories. As such, those that align with cooperative and altruistic beliefs (for example) might be deemed reliable considered judgements to which we should attend, and those that do not might be eliminated. However, if our moral intuitions are supposed to be our access to an independent moral reality, then we can hardly use our moral theories as the yardstick against which to measure their reliability — especially as it is these same moral intuitions that we are supposed to be using as the evidence against which we test the credibility of our moral theories (thereby rendering the whole enterprise circular).

Consequently, I think that some significant portion of the problems for the reliability of moral intuitions caused by interpersonal intrasource inconsistency or disagreement and framing effects probably cannot be satisfactorily resolved by means of a process of reflective equilibrium. Moreover, I think the fact that moral intuitions plausibly result from the conjunction of evolutionary adaptations and distorting cultural factors effectively undermines their input into a process of reflective equilibrium at all. As such, though I concur that there is an independent moral reality, I do not agree that our moral intuitions give us reliable access to that reality.

[79] For arguments to this effect, see: Richard Brandt, *A Theory of the Good and the Right* (New York: Oxford University Press, 1979); Richard Brandt, 'The Science of Man and Wide Reflective Equilibrium', *Ethics*, 100 (1990), 259–78; R.M. Hare, 'Rawls' Theory of Justice', *Philosophical Quarterly*, 23 (1973), 144–55, 241–51.

Perhaps the defender of moral intuitions can take another route here. Moral intuitionists argue that certain moral propositions are *self-evident*, thereby necessitating no additional proof. A self-evident proposition should be distinguished from an *obvious* truth. Self-evidence is a property of a proposition, and so is not relative in the sense that such a proposition could be self-evident to one person but not to another. By contrast, what is obvious to me may not be obvious to you, and so obviousness is relative to particular groups or individuals. Moreover, many obvious truths are not self-evident. For example, it is obvious but not self-evident that the world is bigger than an apple, and that a heavy object will fall if dropped.

Classical intuitionists (e.g. Richard Price and W.D. Ross) argued that some basic moral propositions are self-evident, such that properly understanding the proposition compels assent.[80] This view suffers from the problem that even universal assent about some proposition does not, in and of itself, confer *justification* upon that proposition. Thus, on such a view, a self-evident proposition need not be a justified proposition. However, if is not necessarily justified, then this renders as epistemically dubious the claim that it requires no additional proof. The contemporary intuitionist Robert Audi improves upon this earlier view by claiming that a proper understanding of such propositions *justifies* belief, rather than compelling it. According to him, self-evident propositions are:

> truths such that (a) adequately understanding them is sufficient justification for believing them..., and (b) believing them on the basis of adequately understanding them entails knowing them.[81]

[80] Richard Price, 'A Review of the Principle Questions in Morals', in *The British Moralists 1650–1800, Ii*, ed. by D.D. Raphael (Oxford: Clarendon Press, 1758/1969), pp. 131–98; W.D. Ross, *The Right and the Good* (Oxford: Clarendon Press, 1930/2002).

[81] Robert Audi, 'Intuition, Inference, and Rational Disagreement in Ethics', *Ethical Theory and Moral Practice,* 11 (2008), 475–92 (p. 478).

Observe that, on Audi's account, one may have an adequate understanding of a self-evident proposition, and so have sufficient justification for believing it, yet still not believe it. Accordingly, self-evidence is not a property of a mental state, but instead a property of a proposition.

The claim that certain moral propositions are self-evident is a controversial one (as Philip Stratton-Lake concedes), but can it nonetheless be sustained?[82] If we endorse Audi's conception of what it is to be a self-evident proposition, then I would submit that it cannot. Firstly, we might mount an argument against Audi based upon interpersonal intrasource inconsistency or disagreement. There are many putative self-evident moral propositions that are adequately understood (e.g. by moral philosophers), yet we still find significant and persistent disagreement (amongst moral philosophers, and even intuitionists) as to their veracity. As explained earlier, the correct epistemic response in light of such disagreement will typically be a suspension of belief or an appropriate reduction in credence. However, in that case, those who adequately understand these propositions cannot really be said to *know* them. As such, there are then numerous counterexamples to Audi's general claim, thereby casting doubt upon the view that there are self-evident propositions in his sense. Of course, Audi might respond that there are still *some* putatively self-evident propositions about which there is no disagreement (amongst moral philosophers), and no doubt is thereby cast upon their claim to being self-evident. However, as I think the general claim has been undermined, I would suggest that the onus would be upon Audi to defend some modified version of the claim, applying just to self-evident propositions about which there is no interpersonal intrasource inconsistency or disagreement.

[82] P. Stratton-Lake, 'Self-Evidence, Intuition and Understanding', in *Madison Workshop in Metaethics,* (Madison, WI, 2016), (p. 1).

Secondly, observe, as Stratton-Lake does, that the notion that our understanding a proposition can justify it is a most peculiar one.[83] There are two sources for the peculiarity of this notion. Firstly, epistemic justifications are required to be appropriately linked to truth. With synthetic propositions (both synthetic a posteriori ones and synthetic a priori ones — where the latter is the kind that ethical intuitionists are interested in[84]), the appropriate link to truth of a justifier and the belief it justifies is that the former must constitute *evidence* (understood as something that raises the epistemic probability of the proposition for which it is evidence) for the truth of the latter. However, secondly, our mere understanding of a proposition does not (except in a few self-referential cases) provide evidence for its truth. Hence, our mere understanding of a synthetic (a priori) self-evident moral proposition cannot justify us in believing it, and so even if the proposition (e.g. *p* = there are reasons for the dedicated immoralists to refrain from their evil deeds) was classed as a self-evident moral proposition on Audi's conception of this, understanding it would not justify us in believing it. By contrast, mere understanding may be an appropriate link to the truth in the case of *analytic* propositions, which are true simply in virtue of their meaning.

Note that because all self-evident propositions on Audi's conception are *necessary* (since one could not be justified in believing a contingent proposition simply in virtue of understanding it), but, on Goal Theory, all moral propositions are *contingent*, then, on my account there can be no self-evident moral propositions. Self-evident propositions on Audi's conception would also be knowable non-empirically. Yet, on Goal Theory, what we would desire most, when fully rational and sufficiently informed, as well as what will best serve this desire, are only knowable empirically.

[83] Stratton-Lake, pp. 4-5.

[84] A synthetic a priori proposition I understand as a proposition where the predicate is not logically or analytically contained in the subject (i.e. *synthetic*), and the truth of which is verifiable independently of experience (i.e. *a priori*). Kant and his rationalist followers maintain that there are such propositions (e.g. 'the sum of 7 and 5 is 12', and 'the straight line between two points is the shortest'), but their existence is controversial, with others arguing that such examples are either synthetic a posteriori or analytic a priori.

Thus, once again, there can be no self-evident moral propositions on my account (and likewise on other consequentialist moral theories). As Stratton-Lake points out, some (including Audi) may claim that synthetic a priori truths are *conceptual*. However, as he argues, depending upon our understand of 'conceptual truth', it is either false that synthetic a priori truths are conceptual truths (if we equate 'conceptual truth' with 'analytic truth'); or else the synthetic a priori truths that intuitions are interested in cannot be understood as conceptual truths in the sense at hand (if a conceptual truth is understood as one that anyone with a clear grasp of the relevant concepts would endorse).[85]

So, what could justify us in believing a synthetic self-evident moral proposition (including, perhaps, the proposition in question)? Might our intuition of a self-evident proposition justify us in believing it? Not if we, like Audi, understand intuitions as beliefs of a certain sort (specifically, beliefs that are non-inferred, firmly held, pre-theoretical, and based solely on an understanding of their content), because our belief that *p* cannot justify our belief that *p*. The same applies if intuitions are inclinations to believe, since the fact that I am inclined to believe some *p* is not a justification for believing *p*. Another problem with Audi's account, as noted by Stratton-Lake, is that is fails to capture adequately the *recalcitrance* of intuitions, i.e. the fact that intuitions can remain even when the agent does not believe them. For example, I have an intuition that a bullet dropped from a certain height will hit the ground before one fired horizontally at that height from a gun — even though I know this intuition is false. Yet, on Audi's account, it is not possible to have an intuition that *p* without a belief that *p*. Perhaps this justificatory role might be permitted on some alternative account of intuitions, but what form might this alternative account take?

[85] Stratton-Lake, p. 6.

As mentioned earlier, Shafer-Landau advocates a moral epistemology that incorporates a form of intuitionism.[86] On his account, some moral facts can be known because they are self-evident, and the others can be known because we have reliable methods to discover them. However, I think his account also fails, because it faces a similar challenge to that faced by Audi's, viz. merely understanding and attentively considering a moral principle does not provide *evidence* for its truth. And this undermines his use of supposedly self-evident moral principles from which he thinks we can justify moral judgements arrived at non-inferentially from said principles.

Stratton-Lake prefers Bealer's account of intuitions, according to which intuitions are a mental state that he calls *intellectual seemings*, where these are distinct from beliefs or judgements.[87] Intellectual seemings are analogous to *perceptual* seemings. In the same way that things can seem perceptually to be a certain way, e.g. coloured, then certain propositions can seem to be true, or present themselves to the mind as true. Something can seem true to us, even if we do not believe it, so these seemings are not beliefs. So, does this understanding of intuitions supply the intuitionist with the necessary means to justify us in believing that the propositional content of some moral intuition is self-evident? Well, just as, absent any undercutting defeaters or other outweighing considerations, what supposedly justifies some experiential belief is that things seem, perceptually, to be that way (e.g. my belief that there is a tree outside the window), then, analogously, absent any undercutting defeaters or other outweighing considerations, what would justify my belief in a self-evident proposition is that it seems true. The basic idea is that, unless we have some reason to distrust an intuition, then we may believe that things are the way they seem perceptually, and, analogously, morally. On this understanding, a self-evident

[86] See: Shafer-Landau, *Moral Realism: A Defense*, pp. 229-302.
[87] G. Bealer, 'Intuition and the Autonomy of Philosophy', in *Rethinking Intuition,* ed. by M. DePaul and W. Ramsey (Lanham, Md.: Rowman & Littlefield, 1998), pp. 201-40.

proposition would be one where a clear intuition is sufficient justification for believing it, and for believing it based on that intuition.

I do not wish to enter here into a discussion of perceptual experience and epistemological justification — in particular, whether perceptual seemings can be justifiers.[88] Instead, for this purpose, I shall just assume, along with the intuitionist who posits the above analogy, the view that things are mostly the way they seem perceptually — absent any undercutting defeaters or other outweighing considerations. (Though the assumption that things are mostly the way they seem to be, perceptually, may be unwarranted. In particular, science has found that almost nothing actually correlates with perceptual seeming — e.g. colours do not exist, solid objects do not exist, the sun is not moving, the mind is not disembodied; and that is all before we even get to sensory and cognitive illusions. As such, even the starting point for the intuitionist's case may be fatally undermined.) However, I would argue that perceptual seemings and intellectual seemings are disanalogous in a crucial way that would undermine any reliance upon them by the intuitionist. To see why, consider the general form of the claim being made here:

> **(X)**: if something seems to be a certain way in domain *T*, then, absent any undercutting defeaters or other outweighing considerations, we are justified in believing that it is that way.

Now, imagine that there is a seeming *S* in a domain *T* for which we have no particular undercutting defeaters or outweighing considerations. In that case, the epistemic probability that the content of *S* in domain *T* is true would equate to the *prior*

[88] But for a discussion of that question, asking in particular whether some version of *dogmatism* (i.e. in the absence of defeaters, a [typically] perceptual seeming that *P* provides justification to believe *P*) or *phenomenal conservatism* (i.e. in the absence of defeaters, *any* seeming that *P* provides justification to believe *P*) can be sustained, see: Chris Tucker (ed.), *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism* (Oxford: Oxford University Press, 2013).

probability that the content of *S* in domain *T* is true. To see why, consider the long form of Bayes' Theorem (which is a mathematical formula for calculating conditional probabilities)[89]:

$$P(h|e.b) = \frac{P(h|b) \times P(e|h.b)}{[P(h|b) \times P(e|h.b)] + [P(\sim h|b) \times P(e|\sim h.b)]} \qquad [\mathbf{1}]$$

The relevant terms being:

- $P(h|e.b)$: the epistemic probability that our hypothesis, *h*, is true given the entirety of our reliable background knowledge, *b*, and the reliable evidence directly relevant to what our theories seek to explain, *e* (where *b* + *e* is the entirety of our knowledge).
- $P(h|b)$: the *prior probability* that *h* is true, where this is the unconditional probability that is assigned to *h*, given the entirety of our reliable background knowledge, *b*, and before we take into account any evidence, *e*, that is directly relevant to *h* [NB. $P(\sim h|b)$ is the converse of $P(h|b)$].
- $P(e|h.b)$: the consequent probability of *e*, if *h* is true, given *b*.
- $P(e|\sim h.b)$: the consequent probability of *e*, if *h* is false, given *b*.

In this case, our hypothesis, *h*, is that the content of *S* in domain *T* is true. *Ex hypothesi*, there is no particular evidence for or against *h*. Accordingly, we would assign a neutral value of 0.5 to both the consequent probability of *e*, if *h* is true, given *b*; and to the consequent probability of *e*, if *h* is false, given *b* — i.e. to $P(e|h.b)$ and $P(e|\sim h.b)$ respectively. Substituting these values into [1] gives us:

[89] See: James Joyce, 'Bayes' Theorem', *The Stanford Encyclopedia of Philosophy*, Winter 2016 edn, ed. by Edward N. Zalta <https://plato.stanford.edu/archives/win2016/entries/bayes-theorem/> [accessed 6th March 2019].

$$P(h|e.b) = \frac{P(h|b) \times 0.5}{[P(h|b) \times 0.5] + [P(\sim h|b) \times 0.5]}$$

Then, noting that $P(\sim h|b)$ is the converse of $P(h|b)$, we have:

$$P(h|e.b) = \frac{P(h|b)\ \times 0.5}{0.5}$$

Thence:

$$P(h|e.b) = P(h|b) \qquad [\mathbf{2}]$$

Now, I have assumed for my purposes that things are mostly the way they seem perceptually (absent any undercutting defeaters or other outweighing considerations). As such, when *T* is taken to be the perceptual domain, the prior probability that the content of any particular seeming is true will be high. Given equation [2] above, it follows that the epistemic probability that the content of any particular *S* in the perceptual domain is true — i.e. $P(h|e.b)$ — will be equally high (absent any undercutting defeaters or other outweighing considerations). And if this epistemic probability is high (e.g. >0.95), then we would probably be justified believing that things are the way they seem, perceptually, according to *S*.

By contrast, when *T* is taken to be the moral domain, we find that how propositions seem is routinely not as things actually are. As discussed earlier, even if we set aside the fact that such moral seemings regularly contradict one another, are contradicted by empirical evidence, and vary between people and groups of people, they are frequently vulnerable to framing effects and biases. Moreover, some subset of them are likely caused by unreliable mental processes; and, for evolutionary and cultural reasons, they need not (and probably do not) track moral reality. For all of these reasons, the prior probability that the content of any particular moral seeming is true is probably significantly lower than in the perceptual case. Given equation [2] above, it follows that the epistemic probability that the content of any particular *S* in the moral domain is true will be correspondingly lower. And if this epistemic probability is significantly lower (e.g. <<0.95), then we would probably *not* be justified believing that things are the way they seem, morally, according to *S*, given that justification requires a suitably high epistemic probability.

Thus, I would deny that the claim being made in (X) is true in general. Even if it is true in the perceptual domain (as intuitionists who make this argument assume), I would argue that it is plausibly *false* in the case of the moral domain. As such, I would argue that we cannot infer from the assumption that perceptual seemings are justifiers that moral seemings are thereby justifiers too. In fact, I would argue that, in general, moral seemings are probably *not* justifiers. Accordingly, I would suggest that the following claim is a much more plausible one with regard to moral seemings:

> **(X*)**: if something seems to be a certain way in the moral domain, then we are only justified in believing that it is that way if we have sufficient independent supporting reason or evidence.

In light of this, I would say that arguing the claim that some moral proposition (e.g. *p*, where *p* = there are reasons for the sensible knave to refrain from his or her evil deeds) is *self-evident*, thereby necessitating no additional proof, would not be at all rationally persuasive (where I understand a self-evident proposition to be one where a clear intuition is sufficient justification for believing it, and for believing it on the basis of that intuition). I agree with Stratton-Lake when he says that:

> Once we learn that it is our intuition of some self-evident proposition rather than our understanding of it that justifies us in believing it, we can see that all of the epistemic work is done by moral intuitions. They are the things that do the justifying. We can call a subclass of intuitive propositions self-evident, but once we get clear on what that means, all we are saying is that that proposition is such that an intuition of it justifies us in believing it, and provides a strong enough justification to ground knowledge.[90]

However, because I take moral intuitions to not be reliable indicators of truth (for all of the reasons presented), then I think that they do no useful epistemic work for the intuitionist, thereby undermining the thesis that some moral propositions are self-evident and therefore in need of no further proof. In view of that, then, returning to the idea of moral intuitions as heuristics, I find myself in agreement with Sinnott-Armstrong et al. when they say that:

> …if moral intuitions result from heuristics, moral intuitionists (cf. Stratton-Lake, 2003) must stop claiming direct insight into moral properties. This claim would be as implausible as claiming direct insight into probability or numbers of seven-letter words, based on how we employ the representativeness and availability heuristics. Heuristics often seem like direct insight, but they never really are direct insight, because they substitute attributes. If moral judgments are reached through mediating emotions or affect, then they are not reached directly.[91]

I have spent some time defending premise P2 of Argument 7 (i.e. we probably have [undefeated] reasons to think moral intuitions unreliable), as I think this is the premise most likely to be challenged. And I have also defended P1 (i.e. probably, one is justified in believing on the (sole) basis of a putative source of evidence only if one

[90] Stratton-Lake, p. 22.
[91] Sinnott-Armstrong, Young, and Cushman, 'Moral Intuitions', pp. 267-68.

lacks (undefeated) reason to think it unreliable). I shall not defend the connection in premise P3 between a belief being unjustified and its exclusion from being considered reliable evidence directly relevant to what our theories of normative reasons need to explain — as I think this is probably uncontentious. Thus, I would argue that I have at least *motivated* conclusion C2 of Argument 7, viz. moral intuitions should probably be excluded from being considered reliable evidence directly relevant to what our theories of normative reasons need to explain.

Of course, this is not the end of the matter. I have not produced a knockdown argument against the reliability of moral intuitions. There are always more arguments that can be made, so perhaps my worries can be assuaged. However, I think I have succeeded in the more modest aim of at least motivating the conclusion that moral intuitions are unreliable in general. As such, I would argue that moral facts are not knowable by moral intuition alone (though, as I argued in section 2.3, I think they *are* in principle knowable by empirical investigation, so I am not endorsing moral scepticism). I would call for more research on such issues as the precise psychological nature of moral intuitions, determining the conditions under which epistemic circularity is problematic, and the nature and extent of disagreement about the content of intuitions. In the meantime, I find myself in agreement with Kagan that the appropriate stance to take toward our moral intuitions will involve accepting an error theory, according to which at least many of our case specific moral intuitions are mistaken.[92] However, without determining first from our story about the mechanics of moral intuition precisely how and when our moral intuition fails to track moral reality, it is difficult to be confident that the requisite error theory can be produced.

[92] Kagan, 'Thinking About Cases', p. 49.

3.5 Shafer-Landau redux

Returning now to Shafer-Landau's argument, what is the consequence of the foregoing excursus? Well, I would argue that being reliable in the trustworthy sense is a high but appropriate epistemic standard in this context, and that moral intuition probably fails to meet it. As such, I think we should affirm the general unreliability of moral intuitions, and therefore agree that the proposition in (A*) (i.e. if we all have an *intuition* that *p*, then *p*) is probably unjustified.

The form of the proposition in (A*) excludes an attack on the reliability of moral intuitions based upon interpersonal intrasource inconsistency or disagreement, since the antecedent specifies that *everyone* has the intuition that *p*. However, even if Shafer-Landau's particular intuition really is one about whose propositional content everyone agrees (which is an undefended claim), it is nonetheless the case that, absent independent support, it would not be known to be *reliable*. Moreover, it is still vulnerable to attacks based on framing effects, biases, and the likely failure of moral intuitions to reliably track moral truth. These, I would argue, are sufficient to undermine any claim to its reliability, even if it garners universal agreement.

If (A*) is probably unjustified, then, absent independent support, premise P2 of Argument 6 (in section 3.3), which depends upon its truth, will be unproven and epistemically dubious, leaving conclusion C1 of Argument 6 also unproven. Thus, Shafer-Landau's argument for the falsity of the HTR* would be undermined, leaving my account standing. Perhaps Shafer-landau might be able to reformulate his argument in some stronger form. However, it is difficult to see how he can avoid any dependence upon moral intuitions (and, given his epistemological commitments, he may not wish to do so anyway). In that case, any reformulation would also seem to be undermined if moral intuitions are generally unreliable. Alternatively, he might concede that moral intuition is generally unreliable, but argue that, nonetheless, the particular intuition in question is a trustworthy one; or perhaps argue that the special faculty of intuition whereby we can (supposedly) reflectively access moral facts is somehow distinct from the kinds of moral intuition that I have evaluated here. I will assess such an argument when it is presented. In the meantime, I submit that my account plausibly resists Shafer-Landau's argument.[93]

Returning to Shafer-Landau's sample dedicated immoralists from section 3.3, I would submit that, in the real world, both the enthusiastic torturer and the person who takes delight in failing to rescue the toddler very likely make a serious mistake in acting as they do. Even if their *prima facie* immoral actions do serve their present, unenlightened desires, with these desires not being served at all by refraining, I would suggest (for the kind of reasons adduced in section 2.5) that their *enlightened* desires would be strongly served by refraining. In that case, on the HTR*, they would then have excellent reasons to refrain. However, if Shafer-Landau were to stipulate that even the immoralists' enlightened desires would be served by acting in the way that they do, with them being served not at all by refraining, then I would submit that they would then have *no* reason to refrain, notwithstanding any moral intuition to the contrary. (That we obtain a counter-intuitive result should come as no surprise when we construct a bizarre thought experiment.)

Shafer-Landau sums up his motivation for rejecting instrumentalism when he says that:

> We cannot prove that there are categorical reasons. But when we vividly contemplate a world without them, one in which there is literally no consideration that stands against the actions of a torturer, and none in favour of easily rescuing a child from imminent

[93] Shafer-Landau also presents another argument, based upon considerations of responsibility and blameworthiness, as these relate to fanatics who blow up innocent civilians. For reasons of space, I shall set this argument aside, but I think it is no more successful than his first.

> death, most of us will find that instrumentalism has as much appeal as the various sorts of scepticism that we take seriously only in the study.[94]

However, I would submit that my account eliminates almost all of that motivation for denying instrumentalism (at least, in the form of the HTR*) — since, in the real world, the torturer almost certainly *does* have a reason to refrain from his actions, and the potential rescuer almost certainly *does* have a reason to save the child. At the same time, it does not predicate its case upon what I have argued is epistemically dubious moral intuition. Accordingly, while I think Shafer-Landau has failed in his aim to show that there are categorical reasons, and thereby failed to resist the relativistic conclusion that the content of moral requirements is contingent on our commitments, I think this is unproblematic in practice, because in the real world even hypothetical dedicated immoralists will have excellent reason to align their actions with our stock moral truisms. However, if ever they did not, then we should (provisionally) grant this counter-intuitive result, just as we do with many well-supported but counter-intuitive results in other domains (science is replete with such cases).

3.6 The burden of proof

In light of my findings, it seems that neither Joyce nor Shafer-Landau’s argument succeeds, and so the case for categorical reasons is not made. However, Shafer-Landau claims it is the *instrumentalist* who bears the burden of proof in such a situation. As he says:

> [Instrumentalism] should not, in any event, be regarded as the default view about the nature and source of our practical reasons. If we are ever to accept instrumentalism, we

[94] Shafer-Landau, 'A Defence of Categorical Reasons', p. 205.

> must not only find fault with the arguments that I (and others) have offered on behalf of categorical reasons. We must also be impressed enough to move away from the default position of pluralism about the ultimate sources of practical reasons.[95]

If Shafer-Landau is correct, then, even if his argument for categorical reasons fails, we still ought to adopt as our default position the view that there are categorical reasons (amongst other kinds of reasons). However, I deny it is I who bears the burden of proof here, because I deny that postulating categorical reasons should be our default position in these circumstances. Instead, I would argue that on considerations of ontological parsimony, where entities should not be multiplied beyond necessity, positing categorical reasons in these circumstances would be metaphysically extravagant. I think a little Bayesian analysis will clarify why that is so. Recall the long form of Bayes' Theorem:

$$P(h|e.b) = \frac{P(h|b) \times P(e|h.b)}{[P(h|b) \times P(e|h.b)] + [P(\sim h|b) \times P(e|\sim h.b)]} \qquad [\mathbf{1}]$$

As already noted, the prior probability of some theory is the unconditional probability that is assigned before we take into account any evidence that is directly relevant to that theory. In other words, given just our background knowledge (which is everything we know with reasonable certainty from science, history, and so on), how probable is that theory? In practice, it may be difficult to assign an accurate and objective value to this probability, but it is generally accepted that, *ceteris paribus*, a theory will have a *higher* prior probability to the extent that it is consistent with what we already know to be true (i.e. it is *plausible*), and does not introduce additional elements that are not themselves supported by independent evidence (i.e. it is *ontologically parsimonious*). If a theory

[95] Shafer-Landau, 'A Defence of Categorical Reasons', p. 204.

were *not* consistent with what we already know to be true, then its truth would entail the denial of things that have a very high epistemic probability, leading to a corresponding reduction in its prior probability. And if a theory introduces additional elements that are not themselves supported by independent evidence, then it is committed to things that have a significantly less than 100% probability of existing or obtaining, with the consequent negative impact upon its prior probability.

With respect to the particular case at hand, consider two rival theories of normative reasons that aim to explain the entire collection of relevant evidence, *e* (the precise contents of *e* need not concern us at this stage). The first, h_1, is Goal Theory's implied theory of normative reasons, viz. the HTR*, whereby an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. In postulating only *hypothetical* reasons for action, h_1 is committed to the existence of a type of entity that is already established and uncontentious. The second theory, h_2, is also committed to the existence of hypothetical reasons (including, let us reasonably suppose, those postulated by h_1), but is additionally committed to the existence of a new and unproven type of entity, viz. *categorical* reasons. Thus, the ontological commitments of h_1 form a proper subset of those of h_2.[96] Moreover, h_2's additional commitments are to a kind of entity that has a significantly less than 100% probability of existing. If there is no evidence for or against its existence, then such an entity would necessarily have a 50% probability of existing; and if we have evidence *against* its existence, then this probability would be less than 50%, perhaps substantially so.[97] What can we now say about the relative epistemic probabilities of h_1 and h_2?

[96] Theory h_2 may additionally postulate hypothetical reasons for action that h_1 does not, e.g. reasons that are ancillary to the satisfaction of agents' *non*-true strongest desires. However, that has no bearing upon my argument.

[97] See: Richard Carrier, *Proving History: Bayes's Theorem and the Quest for the Historical Jesus* (Amherst, NY: Prometheus Books, 2012), pp. 41-96.

Well, since h_1 is more ontologically parsimonious than h_2, then h_2 incurs a reduction in its prior probability relative to h_1. As a result, *ceteris paribus*, h_1 will have a higher prior probability than h_2. That is, the value of $P(h_1|b)$ will, *ceteris paribus*, be higher than $P(h_2|b)$. Moreover, in terms of plausibility, a theory that introduces new and unproven elements can only ever be equally or less consistent with what we already know to be true than a theory that does not do this, *ceteris paribus*. Thus, *ceteris paribus*, considerations of plausibility will either leave this disparity between $P(h_1|b)$ and $P(h_2|b)$ unchanged, or they will increase it (but never *decrease* it).

Is this analysis undermined by the so-called problem of subjective priors, whereby the assignment of the priors is deemed subjective, and, as such, not representing objective reasoning? No, because I am discussing only *relative* priors, not *absolute* ones. And from my preceding analysis, it is not subjective that, whatever their absolute values, *ceteris paribus,* we have:

$$P(h_1|b) > P(h_2|b) \quad [2]$$

That is all I require here (making a standard assumption that a theory's prior probability is composed solely of its plausibility and parsimony).[98]

To see what effect this has upon the relative epistemic probabilities, I must now incorporate the *consequent* probabilities. At this point, Shafer-Landau might want to argue that h_2 achieves better evidential fit than h_1 does, with categorical reasons pulling their weight in explanatory theories, such that they are required in our best overall explanatory picture of the world. However, given the apparent lack of success of his

[98] For reasons to think that the problem of subjective priors does not trouble Bayes' Theorem more generally, see: Carrier, *Proving History: Bayes's Theorem and the Quest for the Historical Jesus*, p. 81. See also: Giulio D'Agostini, 'Role and Meaning of Subjective Probability: Some Comments on Common Misconceptions', in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering,* ed. by Ali Mohammad-Djafari (Melville, NY: American Institute of Physics, 2001), pp. 23-30.

argument for categorical reasons, I would submit that this is precisely what Shafer-Landau has so far *failed* to demonstrate. As such, I would submit that h_2 does not improve upon h_1 in terms of evidential fit. Accordingly, until a more cogent argument is for categorical reasons is presented, I shall make the assumption that both theories fit the evidence equally well, taking $P(e|h_1.b) = 1$, $P(e|h_2.b) = 1$, $P(e|{\sim}h_1.b) = 1$, and $P(e|{\sim}h_2.b) = 1$. (I could have chosen any probability values between 0 and 1 here, so long as they were the same for each theory, but choosing 1 simplifies the subsequent calculations.) In that case, substituting into equation [1] we get:

$$P(h_1|e.b) = \frac{P(h_1|b) \times 1}{[P(h_1|b) \times 1] + [P({\sim}h_1|b) \times 1]}$$

And noting that $P({\sim}h|b)$ is the converse of $P(h|b)$, we have:

$$P(h_1|e.b) = \frac{P(h_1|b)}{P(h_1|b) + [1 - P(h_1|b)]}$$

That is,

$$P(h_1|e.b) = P(h_1|b) \quad [3]$$

Going through the same steps for h_2, we would get:

$$P(h_2|e.b) = P(h_2|b) \quad [4]$$

Combining equations [3] and [4] with equation [2], we get:

$$P(h_1|e.b) > P(h_2|e.b) \quad [5]$$

That is, *ceteris paribus, h_1* will have a higher epistemic probability than h_2. Therefore, *ceteris paribus,* it would be rational to prefer h_1 to h_2.

I submit that this result places the burden of proof onto the proponent of h_2 if they wish to argue that we ought to prefer their theory to h_1. That is, I think our default position ought, *ceteris paribus*, to be the more epistemically probable (and thence rational) one, viz. h_1. It would be bizarre to claim that it ought to be the *less* epistemically probable (and thence irrational) one, as Shafer-Landau's view would suggest. Thus, where the case for categorical reasons is not made (as is the situation here), then I would argue that our default position should be to exclude them from our ontology, not to affirm them as part of a pluralist view about the ultimate source of normative reasons, as Shafer-Landau wants to do.

3.7 Conclusions

The aim of this chapter was to respond to the challenge that there are categorical normative reasons, contra Goal Theory's Humean account. This I have done, arguing that Goal Theory successfully resists the challenge.

Firstly, I evaluated the Central Problem, and undergeneration and overgeneration objections (all of which are thought to be problematic for the HTR), finding that Goal Theory's implied theory of normative reasons (the HTR*) offers a plausible response to these. After that, I evaluated two representative arguments that the proponent of categorical reasons might advance. The first, from Richard Joyce, intended to demonstrate, by reference to the example of the so-called sensible knave, that categorical reasons are a non-negotiable element of morality (though he then denies that there are such reasons, thereby denying moral reality). By contrast, the second, from Shafer-Landau, argued by reference to the hypothesised dedicated immoralist that there are categorical reasons (though, contra Joyce, he denies that categorical reasons are a non-negotiable element of morality). I found that neither argument succeeded, meaning that the case for categorical reasons was not made.

I then evaluated Shafer-Landau's claim that, in the absence of a successful argument for categorical reasons, our default position should be a pluralist one with regard to the possible sources of normative reasons, thereby admitting categorical reasons into our ontology nonetheless. By means of a Bayesian analysis, I countered this claim, concluding that our default position should instead be to exclude categorical reasons. Accordingly, I would argue that my account plausibly survives the challenge that there are categorical normative reasons (at least, as represented by the arguments in question).

In a pattern that we will see repeated over the next few chapters, I would suggest that while my account gives up on categorical reasons, on the basis that they are not defensible, it still captures *enough* of what is insisted upon by those proposing categorical reasons to be, overall, plausible. Specifically, my account delivers the *authority* that Joyce insists upon, as well as *inescapability**, where I think the conjunction of the two is all that that can be reasonably demanded, and is sufficient to undercut the motivation for inescapability (since moral evaluations still apply to agents, or their actions, independently of agents' *present* desires, which I think is what the relevant intuition is getting at). Yet, crucially, this conjunction does not then entail categorical reasons. With regard to what Shafer-Landau insists upon, my account accommodates the intuition that there are reasons for dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their *unenlightened* desires.

However, if Shafer-Landau wants to make the further claim that there are reasons for dedicated immoralists to refrain from their evil deeds, despite refraining serving none of their *enlightened* desires, then I would demur, arguing that any corresponding moral intuition is probably unreliable.

Thus, I suggest that my account improves both upon other variants of the HTR, which struggle to deal with the Central Problem and with undergeneration and overgeneration objections, and upon accounts that posit categorical reasons, where I have argued that such entities are metaphysically extravagant.

Having now responded to the first of the three dominant challenges to accounts of Goal Theory’s type, in the next chapter I shall evaluate the second, viz. that normative facts and properties are ‘just too different’ from natural facts and properties to be reducible or identical to them, contra Goal Theory’s naturalistic account of normativity.

Chapter 4

The just too different objection to normative naturalism

How is the moral world related to the non-moral world? Some relationship or dependence must exist. But what is its nature? Whatever the answer, is it not intuitively obvious that the normative is *just too different* from the natural for the former to be a subset of the latter? In that case, if we adopt a naturalistic worldview, then we may end up denying that there are normative facts (as anti-realists do); and if we affirm normative facts, then we may end up denying that the world is completely natural (as non-naturalist realists do). Either way, is not any account that identifies normative and natural facts clearly defective? Yet what if we are wrong about this, with our intuition here being no more reliable than the intuition that caterpillars are just too different from butterflies to be the same species?

In this chapter, I aim to answer the challenge that normative facts and properties are *just too different* from natural ones to be reducible or identical to them, contra Goal Theory's ethical naturalist account. I seek to anchor normativity in the natural world, arguing that normative facts and properties are nothing over and above natural facts and properties. To that end, I shall explain the 'just too different' intuition, present a positive argument for Goal Theory's variety of naturalistic normativity, and evaluate two variations of the just too different (JTD) objection to normative naturalism. My intention is to show that Goal Theory resists both. Here I shall focus upon normativity rather than evaluation. This is not because I think that there is no distinction between the two, or that the latter is subservient to the former. Instead — rightly or wrongly — this is where the debate is ordinarily located, so, for reasons of simplicity and economy, I shall concentrate my efforts here too.

4.1 The 'just too different' intuition

Can Goal Theory survive the most dominant objection to naturalist theories, viz. the JTD objection to normative naturalism — according to which normative facts are *just too different* from natural facts to be reducible or identical to them? In what follows, I shall argue that it can. In so doing, I mean to show how my account meets FitzPatrick's requirement:

> …that such things as goodness and badness, rightness and wrongness, and reasons for acting can all be captured entirely within a metaphysically naturalistic worldview — a conception of reality as containing only the sorts of entities and properties that are either susceptible to investigation by the empirical sciences or at least fully constructible from those that are. Such a claim thus stands in opposition both to the nihilistic denial of normativity altogether and to the nonnaturalist's insistence that normativity is real but can be captured only within a partly nonnaturalistic framework.[1]

David Enoch identifies the JTD intuition, and thinks that it is the underlying motivation for non-naturalism.[2] As he says:

> [N]ormative facts and properties . . . are just too different from natural ones to be a subset of them.[3]

[1] W. FitzPatrick, 'Skepticism About Naturalizing Normativity: In Defense of Ethical Nonnaturalism', *Res Philosophica,* (2014), 559–88 (p. 560).
[2] Enoch, *Taking Morality Seriously,* p. 80.

Enoch supposes this is a pre-theoretical thought that we have prior to any metaethical theorizing. Other non-naturalists share Enoch's intuition. For example, Jonathan Dancy expresses a similar sentiment when he says that:

> There remains a stubborn feeling that facts about what is right or wrong, what is good or bad, and what we have reason to do have something distinctive in common, and that this common feature [normativity] is something that a natural fact could not have.[4]

In short, many non-naturalists claim that normative (including moral) facts and properties intuitively feel *just too different* from natural facts and properties to be of the same kind. As such, they claim, any naturalistic theory (including Goal Theory) that posits an identity between normative facts and properties and natural ones is thereby defective.

On its own, I think the JTD intuition identified by Enoch bears little weight. After all, I argued in section 3.4 that moral intuitions are unreliable in general, that we have no generally accepted means to distinguish any trustworthy intuitions from untrustworthy ones, and so beliefs based (solely) on such intuitions are probably not justified. As such, absent independent reason to think that the JTD intuition is a trustworthy one, I think we should adopt a position of scepticism.

[3] Enoch, *Taking Morality Seriously*, p. 100.

[4] J. Dancy, 'Nonnaturalism', in *The Oxford Handbook of Ethical Theory*, ed. by D. Copp (Oxford: Oxford University Press, 2005), (p. 136). For expressions of similar intuitions, see: A. Donagan, 'W. A. Frankena and G. E. Moore's Metaethics', *Monist*, (1981), 293-304; W. J. Fitzpatrick, 'Robust Ethical Realism, Non-Naturalism and Normativity', *Oxford Studies in Metaethics*, (2008), 159-206; W. J. Fitzpatrick, 'Ethical Non-Naturalism and Normative Properties', in *New Waves in Metaethics*, ed. by M. Brady (Basingstoke: Palgrave Macmillan, 2010), pp. 7-35; M. Johnston, 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society (Supp.)*, (1989), 139-74; C. McGinn, *Ethics, Evil, and Fiction* (Oxford: Clarendon Press, 1997); Scanlon, *What We Owe to Each Other*.

So, does Enoch adduce any independent reason to think that the intuition is a trustworthy one? Well, at the end of a discussion in which he seeks a positive argument supporting his irreducibility claim (conceding that he does ‘not have such an argument up my sleeve’, and saying that ‘there is some reason to think that we find ourselves here in a dialectical predicament where no such argument is possible’), Enoch admits that we are left:

> Where we started, I'm afraid — with the just-too-different intuition. Normative facts sure seem different from natural ones, different enough to justify an initial suspicion regarding reductionist attempts … We may not be able to do here much more than just stare at the just-too-different intuition and try to see how plausible it seems to us. … And to me, it seems very plausible indeed.[5]

Thus, we seem to be left with little more than a bare intuition; and I think we ought to be wary of an argument that depends upon the undefended intuition that two things are just too different from one another to be considered to be of the same kind, when the precise nature of the difference cannot be explained. Caterpillars intuitively seem just too different from butterflies to be the same species, and yet they are; and heat intuitively seems just too different from molecular kinetic energy to be the same thing, yet it is. Moreover, I am not merely adducing some opposing intuition on which normative facts and properties seem *not* too different from natural ones for the former to be a subset of the latter. Rather, I have presented what I take to be a cogent (albeit defeasible) argument on which this is implied, viz. Argument 1 from section 2.1.

I shall return to the JTD objection shortly, where I shall sharpen it into two main variants that may be pressed against Goal Theory’s account. However, before I

[5] Enoch, *Taking Morality Seriously*, p. 108.

do that, let me first adduce a positive argument for the particular version of naturalistic normativity that I am defending.

4.2 A positive argument for naturalistic normativity

I think that normative facts and properties exist, and are a subset of the natural ones. However, beyond gesturing to Argument 1, how might I make a positive argument for the claim that normativity is natural, contrary to the objection that normative and natural facts seem just too different to be identified with one another? In essence, I shall argue that all normative concepts are analysable in terms of one, fundamental normative concept; and this putative fundamental normative concept picks out a natural property.

Firstly, with regard to the fundamental normative concept in terms of which I think all normative concepts may be analysed, I shall adopt the popular 'buck-passing' account of normativity, on which normative concepts can be analysed in terms of the concept of a *reason*.[6] As such, I shall concur with (the non-naturalist) Parfit, who thinks that 'normativity is best understood as involving reasons or apparent reasons.'[7] Parfit says that 'when I call some claim normative in the reason-implying sense, I mean roughly that this claim asserts or implies that we or others do or might have some reason or apparent reason.'[8] He thinks that words such as 'good', 'bad', and 'ought' all have reason-implying uses.[9] In addition, he thinks that a fact gives us

[6] E.g. Scanlon, *What We Owe to Each Other*.
[7] Parfit, *On What Matters Vol 2*, p. 269. Copp and Dancy criticise Parfit's reason-implying conception of normativity, but I shall bracket those criticisms here: Copp, 'Normativity and Reasons: Five Arguments from Parfit against Normative Naturalism', p. 35. Also: Dancy, 'Nonnaturalism', p. 136.
[8] Parfit, *On What Matters Vol 2*, p. 268.
[9] Parfit, *On What Matters Vol 1*, p. 33.

reason for something when it ‘counts in favour’ of that thing.[10] Parfit says that ‘all reasons have normative force.’[11] He also specifies that a reason for acting will count as ‘decisive’ if we have most reason to act in one way rather than any other, and when our reasons for doing something are decisive in this way, then this is what we ought to do in the ‘decisive-reason-implying sense.’[12]

As to whether reasons pick out natural properties, remember that on my account of normative reasons, viz. the HTR*, an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. Thus, in that case, being a reason is a *natural* property — on the basis that the desires of fully rational and sufficiently informed versions of agents are natural (as I argued in section 2.2).

Thus, if I am right that all normative concepts are analysable in terms of *reasons*, and that being a reason is a natural property, then it follows that normativity is natural. Moreover, given that, on my account, a normative fact is a fact regarding what one ought to do, and what one ought to do is what one has decisive reason to do (per Parfit’s platitudes), then it follows on my account that normative facts are analysable in terms of reasons. However, since reasons pick out natural properties, then it follows that normative facts are natural facts on my account.[13] Thus, I would submit that the HTR* can explain all normative claims, without error.

Syllogistically, we may say:

[10] Parfit, *On What Matters Vol 1*, p. 31.
[11] Parfit, *On What Matters Vol 1*, p. 35.
[12] Parfit, *On What Matters Vol 1*, p. 33.
[13] Schroeder adopts the same basic strategy, but I think that my variant of this improves upon his, insofar as Schroeder uses hypotheticalism as his theory of reasons, whereas I use the HTR*, and I have argued (in section 3.1) that the HTR* improves upon hypotheticalism. See: M. Schroeder, 'Realism and Reduction: The Quest for Robustness', *Philosophers’ Imprint,* 5 (2005), 1-18; Schroeder, *Slaves of the Passions*.

Argument 8

P1)	If all normative concepts are analysable in terms of one, fundamental normative concept, and if that fundamental normative concept picks out a natural property, then normativity is natural.
P2)	All normative concepts are analysable in terms of one, fundamental normative concept, viz. *reasons*. [On Parfit's reason-implying conception of normativity]
P3)	If an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent, then being a reason is a *natural* property.
P4)	An agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. [The HTR*]
C1)	Therefore, being a reason is a natural property. [P3, P4]
C2)	Therefore, normativity is natural. [P1, P2, C1]
P5)	If a normative fact is a fact regarding what one ought to do, and what one ought to do is what one has decisive reason to do, then normative facts are analysable in terms of reasons.
P6)	A normative fact is a fact regarding what one ought to do.
P7)	What one ought to do is what one has decisive reason to do.
C3)	Therefore, normative facts are analysable in terms of reasons. [P5, P6, P7]

C4)	Therefore, normative facts are natural facts. [C1, C3]

4.3 Variations of the just too different objection

Having described my view of normative facts and properties as natural facts and properties — with normative facts and properties existing and being reductively identified with natural ones — this view becomes a legitimate target for variations of the JTD objection. To see how this objection might work in practice, imagine that *M* is some target normative property, such as being *morally right* for *P* in circumstances *C*, and *N* is a natural property, such as the property that actions have when they best serve *P*'s true strongest desire in *C*. I would then claim that *M* is identical to *N* (despite them not being synonymous), with them picking out (instantiations of) the same properties (just as 'water' picks out the same stuff as H_2O, despite the terms not being synonymous). If *M* is identical to *N*, then the normative fact that *x* is *M* (e.g. giving to charity has the property of being morally right for *P* in *C*) is also identical to the natural fact that *x* is *N* (e.g. giving to charity has the property of best serving *P*'s true strongest desire in *C*).

The JTD objection to the aforementioned view may then be expressed as follows[14]:

[14] Hille Paakkunainen makes a similar argument: Hille Paakkunainen, 'The "Just Too Different" Objection to Normative Naturalism', *Philosophy Compass*, (forthcoming) (p. 4).

Argument 9

P1)	If the normative fact that *x* is *M* were identical to the natural fact that *x* is *N*, then the normative fact that *x* is *M* would lack feature *F*.
P2)	The normative fact that *x* is *M* has feature *F*.
P3)	So, the normative fact that *x* is *M* is not identical to the natural fact that *x* is *N*.
C)	So, *M* is not identical to *N*.

Paakkunainen goes on to identify two main versions of the JTD objection in the contemporary literature: one based upon missing *structure* and the other based upon missing *normativity*. In the first case, *F* would be some structural feature of normative facts or their relation to natural facts. In the second case, *F* would be the normative importance of the target normative facts. I shall now critically evaluate each version, endeavouring to establish if either threatens my account.

4.4 Missing structure

One version of the missing structure variant of the JTD objection depends on the idea that although a natural fact can be the reason for some normative fact obtaining, there is a distinction between a fact’s being normatively relevant in this way and a fact’s being normative.[15] Any normative fact holds *because* some natural fact obtains. For example, if murder is wrong, then there are certain facts about murder that make it wrong. According to this argument, we could not make sense of this if the normative fact *were* that natural fact.[16] Following Copp, we might call this the 'argument from because'.[17]

McNaughton and Rawling explain the supposed problem.[18] Reductive naturalists have to propose a theory that identifies each normative property, *M*, with some particular natural property, *N*. However, in that case, to be at all plausible, it must at least be the case that when something is *M*, it is *M because* it is *N*. Yet if that is so, then a thing's being *M* is not identical to its being *N*. At best, *M* will be identified with some *N* that is merely normatively relevant. The 'because relation' in question is asymmetrical — that is, it cannot hold between a fact and itself.[19] For example, on utilitarian naturalism, the property of moral rightness is the property of maximizing happiness. As such, we are right to do what would maximise happiness. However, according to McNaughton and Rawling, for this to be at all plausible, if an act would be the right thing to do, it is right *because* it would maximise happiness.[20] However, in that case, the theory would be false, because the property of being the right thing to do and the property of being the act we can do that would maximise happiness are distinct properties.

[15] Another version claims that normative facts differ from the natural in always involving reference to *standards*. See: FitzPatrick, 'Skepticism About Naturalizing Normativity: In Defense of Ethical Nonnaturalism'. I shall set this version aside (though I agree with Paakkunainen that it fails).

[16] E.g. Parfit, *On What Matters Vol 2,* pp. 298-303. Here I shall focus in particular upon Parfit's presentation of the argument.

[17] Copp, 'Normativity and Reasons: Five Arguments from Parfit against Normative Naturalism', pp. 43-45.

[18] D. McNaughton and P. Rawling, 'Naturalism and Normativity', *Supplement to the Proceedings of the Aristotelian Society,* (2003), 23–45.

[19] McNaughton and Rawling, 'Naturalism and Normativity', p. 33.

[20] McNaughton and Rawling, 'Naturalism and Normativity', p. 42.

Where does this leave my naturalistic account of normative facts? Well, when I say, for example, that *x* has the property of being *right* for agent *P* in circumstances *C*, *because* doing *x* in *C* has the property of best serving *P*'s true strongest desire, then I am saying that the *same* property is being referred to in two different ways. That is, best serving *P*'s true strongest desire is the *same* as the act being *right* for *P* — this is *what it is* for the act to be right for *P* in those circumstances. What I am *not* saying is that there is some substantive normative sense in which there is a property of *P* doing *x* in *C* that *makes* doing *x* right for them.

I view my proposed identity between normative and natural facts and properties as being relevantly analogous to the identity between heat and molecular kinetic energy, or between water and H_2O, insofar as all are reductive identities between familiar entities (i.e. normative facts and properties, heat, and water) and complex and relatively unfamiliar (albeit natural) ones (i.e. facts about and properties of what will best serve agents' true strongest desires, molecular kinetic energy, and H_2O, respectively), with the reductions in question being unintuitive because there is no direct conceptual connection between the entities being identified. Just as we can correctly identify *explanandum* with *explanans* in these cases (e.g. *X* is water *because* *X* is composed of H_2O), then I submit that we can do the same with normativity (e.g. *X* is wrong for *P because* doing *X* frustrates *P*'s true strongest desire). By contrast, proponents of the argument from because would have to argue that, unlike in the water case, 'because' explanations in the normative case cannot legitimately identify *explanandum* with *explanans*. In order to develop this further, let me define the following identities:

> **(A)**: each heat property, *H*, is identical to some particular property of molecular kinetic energy, *K*
>
> **(B)**: each normative property, *M*, is identical to some particular property of best serving an agent's true strongest desire, *N*

Hence, for example, one might say in the first case that the property of being *hot* is identical to the property of having high molecular kinetic energy. And in the second case, one might say that an action having the property of being *right* for an agent in such and such circumstances is identical to the action having the property of best serving that agent's true strongest desire in those circumstances. Moreover, any use of 'because' would be what we might call a 'reductive' use, rather than a 'causal' one.[21] For example, the object is hotter *because* it has greater molecular kinetic energy (in a reductive, not a causal sense); or the act is right for the agent *because* it will best serve their true strongest desire (again, in a reductive, not a causal sense).

Now, those advancing the argument from because would accept identity (A), and would not consider it a legitimate target for their argument. For example, Parfit says it is a truth that:

> …molecules in some physical object move more energetically, that makes this thing hotter in the pre-scientific sense. Having such greater energy does not cause this thing to be hotter, but is the same as being hotter, or is *what it is* to be hotter. Heat *is* molecular kinetic energy.[22]

[21] Copp refers to this distinction: Copp, 'Normativity and Reasons: Five Arguments from Parfit against Normative Naturalism', p. 44.

[22] Parfit, *On What Matters Vol 2*, p. 299.

On identity (A), each heat property, *H*, is identified with some particular property of molecular kinetic energy, *K*. However, following McNaughton and Rawling's logic, to be at all plausible, it must then at least be the case that when something is *H*, it is *H* *because* it is *K*. Yet nobody would then challenge identity (A) by arguing (as McNaughton and Rawling do about the proposed identity between normative and natural properties) that if that is so, then a thing's being *H* is not identical to its being *K*. Such an argument would clearly miss the target.

However, some would then deny identity (B), and would consider it a legitimate target for the argument from because. Parfit, for example, says of *making right* and *being right* that:

> If there is only a single natural property that makes acts right, we could claim that, when some act has this property, that is the same as this act's being right, or is what it is for this act to be right.[23]

But he then says that the 'claims are, I believe, seriously mistaken', because, in the case of greater molecular kinetic energy making something hotter, this relation 'hold[s] between some property referred to in one way and the same property referred to in another way', but '[t]hat is not true of the relation of making right.' Specifically:

> …there is a trivial sense in which rightness is the property that makes acts right. This is like the sense in which redness is the property that makes things red, and legality is the property that makes acts legal. It is in a different and highly important sense that, when some act has some other property—such as that of saving someone's life—this fact can make this act right. Being an act that saves someone's life couldn't be the same as being right. Nor, I believe, could it be one of the properties in which the

[23] Parfit, *On What Matters Vol 2*, p. 299.

> rightness of acts consists. When some property of an act makes this act right, this relation holds between two quite different properties.[24]

Observe, however, that in terms of identity (B), I do not claim that the fact that doing *x* saved someone's life is what *made* doing *x* right for *P* in *C* (except in a quasi-causal sense, like saying that being on fire is what *made* the object hot). Rather, I want to claim that the action of saving someone's life has the property of being right for the agent because (in a reductive sense) it has the property of best serving that agent's true strongest desire (just as I might claim that the object has the property of being hot because it has the property of high molecular kinetic energy). In that case, the act best serving the agent's true strongest desire is the *same* as the act being *right* for them — this is *what it is* for the act to be right for the agent in those circumstances. As such, I would claim that, on identity (B), the relation would hold between some property referred to in one way and the *same* property referred to in another way. To challenge my proposed identity with the example of saving someone's life would be analogous to challenging the identity of heat and molecular kinetic energy with the example of being hot and being on fire. In both cases, the challenge misses its target, and for the same reason (i.e. in neither case is the proposed identity between the entities being ruled non-identical based on the argument from because).

We do, of course, routinely employ the 'because' relation in a 'making', non-reductive sense. For example, we might use 'it was wrong because it hurt' in this sense. However, as with the saving someone's life example discussed a moment ago, I think this is analogous to saying 'it was hot because it was on the fire'. In both cases, there is what we might call a quasi-causal relationship. Specifically, in the latter case, it is putting the object on the fire that increased its molecular kinetic energy, and the

[24] Parfit, *On What Matters Vol 2*, p. 300.

object is then hotter *because* it has greater molecular kinetic energy (in a reductive, not a causal sense). Likewise, it is that the action hurt that frustrated the person's true strongest desire, and the action is then wrong for the person concerned *because* it frustrated their true strongest desire. The identity (B) is no more troubled by such cases than the identity (A) is troubled by analogous cases.

Parfit goes on to say that the pre-scientific meaning of the word 'heat' (i.e. the property, *whichever it is*, that can have certain effects, such as those of melting solids, turning liquids into gases, causing us to have certain kinds of sensation, etc.) has an explicit gap waiting to be filled — with the concept referring to some property without telling us what this property is. Parfit says that similar claims then apply to the concept expressed by the phrase: 'the natural property, *whichever it is*, that makes acts right.' (Here I would argue that I *have* filled that gap.) Parfit, however, thinks that with regard to the concepts of right and wrong, there is *no* explicit gap waiting to be filled in ways that would allow these concepts to refer to one or more natural properties. As an example, Parfit adduces the concept expressed by the word 'blameworthy'. He claims that:

> …this concept does not refer to some property indirectly, as the property of which something else is true. This concept refers directly to the property of being blameworthy… Though social scientists can discover facts about which are the acts that various people judge to be blameworthy, these are not, I believe, facts about the blameworthiness of these acts. [25]

However, I think on my account that would be false. For example, I might define blameworthiness along the following lines:

[25] Parfit, *On What Matters Vol 2*, p. 302.

> **(C)**: an act *x* done by agent *A* in moral circumstances *C* has the property of *blameworthiness* just in case *A* did *x* in *C* knowing (or being reasonably expected to know) that doing this would not best serve their true strongest desire, and that this would make the act morally wrong.

Of course, because my identity is synthetic, agent *A* does not need to have any thoughts explicitly identifying their true strongest desire to be thinking about what is, in fact, the satisfaction or frustration of such a desire. However, I would argue that, in order to be an appropriate recipient of moral blameworthiness, agent *A* must be cognisant of the fact that they acted against their *true strongest desire*, as opposed to merely knowing that they acted against their deep and abiding satisfaction (or whatever is their true strongest desire), without necessarily knowing that the latter is identical to the former (or necessarily knowing what a true strongest desire even is).[26]

Now, (C) might be mistaken, but to say this without independent argument to support it, on the basis that blameworthiness 'does not refer to some property indirectly, as the property of which something else is true', instead referring 'directly to the property of being blameworthy', would be question-begging against the naturalist. Yet no such independent argument is adduced.

In summary, I would argue that an act being *right* for an agent in such and such circumstances *just is* it being what would best serve that agent's true strongest desire in those circumstances (as water *just is* H_2O, and heat *just is* molecular kinetic energy), and so on for other normative properties. In other words, an act is *right* for an agent in such and such circumstances *because* it best serves the agent's true strongest desire in

[26] On the 'rule-egoism' that I shall discuss in section 5.4, we might formulate a more pragmatic version of definition (C), whereby an act *x* done by agent *A* in moral circumstances *C* has the property of *blameworthiness* just in case *A* did *x* in *C* knowing (or being reasonably expected to know) that this would break a moral 'rule' on Goal Theory, without *A* knowing that they in their particular circumstances constituted a genuine exception to the rule.

those circumstances (with the same *reductive* use of ‘because’ that we find in claims that something is water *because* it is H_2O, or something is hot *because* it has high molecular kinetic energy). The properties and facts involved are self-identical (as with all properties and facts), yet asymmetric (e.g. being hot consists in having a high molecular kinetic energy, but not vice versa). And just as the argument from because would miss its target if aimed at the identities between water and H_2O, and heat and molecular kinetic energy, then, for the same reasons, I suggest that it also misses its target if aimed at my account (even if it might threaten some other naturalist accounts).

4.5 Missing normativity

Let me begin this section with an argument from Parfit. His argument does not obviously fit the outline I gave earlier, insofar as it does not explicitly claim that the normative importance of the target normative facts and properties goes missing in attempts to naturalise them. Nonetheless, Parfit thinks the objection relates to normativity, calling it the ‘Normativity Objection’ to naturalistic accounts.[27] For that reason, I shall evaluate it here.

Any reductive naturalistic theory must identify each normative fact, *M*, with some natural fact, *N*. However, according to Parfit, no natural fact can be normative in his reason-implying sense. Parfit thinks that ‘There is a deep distinction … between all natural facts and such reason-involving normative facts.’[28] Specifically, he says that:

[27] Parfit, *On What Matters Vol 2*, pp. 324-27.
[28] Parfit, *On What Matters Vol 2*, p. 310.

> (A) normative and natural facts are in two quite different, non-overlapping categories.[29]

As such, according to Parfit, naturalism involves a conceptual confusion, since natural facts cannot be normative.

Parfit says that even those defending synthetic forms of naturalism must recognise that the normative concepts constrain what a normative property could possibly be. Though the concepts leave open various possibilities, which must be decided among on non-conceptual grounds, '[m]any other possibilities are, however, conceptually excluded.' This is just as with the concept of heat, where the concept constrains what it could possibly be. Whilst a debate between those who held heat to be molecular kinetic energy and those who held it to be a sort of basic substance such as caloric fluid might have made sense, it 'could not have turned out to be a shade of blue, or a medieval king … given the meaning of these claims, they could not possibly be true.'[30]

Likewise, according to Parfit, ethical naturalism and normative naturalism could not possibly be true. As he says:

> Suppose that you are in the top storey of your hotel, and you are terrified of heights. You know that, unless you jump, you will soon be overcome by smoke. You might then believe … that you have decisive reasons to jump, that you should, ought to, and must jump, and that if you don't jump you would be making a terrible mistake. If these normative beliefs were true, these truths could not possibly be the same as, or consist in, some merely natural fact, such as [the fact that jumping would do most to fulfil your present fully informed desires].[31]

[29] Parfit, *On What Matters Vol 2*, p. 324.
[30] Parfit, *On What Matters Vol 2*, p. 325.
[31] Parfit, *On What Matters Vol 2*, pp. 326-27. The statement in square brackets represents Parfit's proposition (C), which he refers to from within the passage quoted.

Parfit is saying that for one to claim that the property of being what you ought to do is identical to the property of being what will fulfil your present fully informed desires is like claiming that heat is identical to a medieval king. However, just as the latter is conceptually excluded, then so is the former — or so he claims.

Accordingly, Parfit appears to be claiming that the naturalist who would identify normative and natural facts is making an obvious *category mistake* — just as, for example, with the person who would identify heat with a medieval king (or a shade of blue). Does this claim withstand critical scrutiny? Certainly, on the face of it, there does seem to be an obvious category mistake in the case of heat and medieval kings, insofar as making such an identity would seem to involve a fundamental misunderstanding of the nature of the things being talked about. However, observe that nobody who is not suffering from some severe cognitive defect would genuinely affirm such an identity. Yet many metaethicists (who one would hope are not similarly afflicted) *do* affirm an identity between normative and natural facts. As such, if there were a category mistake here, then it would seem to be a rather less obvious one than is involved in identifying heat with medieval kings. As such, I think Parfit's claim is already under pressure. However, if one claims that, nonetheless, identifying normative and natural facts is a category mistake, then can the nature of this mistake be located?

In order to determine this, let me first define some terms: let *x* = natural facts and *y* = normative facts. Now, what is a *fact*? Well, a fact I understand as being a true proposition. (Here I set aside the debate over competing theories of truth, and will just understand truth as being the quality of those propositions that accord with reality.) With regard to *natural*, I shall use the definition that I discussed in section 2.2, whereby a fact is natural if it part of the subject matter of our current natural and social science. Thus, by a *natural fact*, I shall mean a true proposition that is part of the subject matter of our current natural and social science. How would I define a *normative* fact? Well, on Parfit's own reason-implying definition, a normative fact would be a fact that asserts or implies that we or others do or might have some reason or apparent reason to do something.[32] So, when *x* = natural fact, and *y* = normative fact, what are we to make of the claim that *x* and *y* are clearly in different, non-overlapping categories?[33]

Let me consider putative systems of ontological categories. On the system devised by Joshua Hoffman and Gary Rosenkrantz, *x* and *y* appear in the *same* basic category, viz. *Proposition*.[34] Moreover, for Reinhardt Grossman, they would once again be located within the same basic category, viz. *Facts*.[35] I can find no system on which they would be in *different* basic categories.

However, even if he were to concede an apparent failure to distinguish between the categories containing certain natural and normative facts, Parfit (and other non-naturalists) might argue that any failure to demonstrate that natural and normative facts are in different basic categories exposes a deficiency in our candidate systems of ontological categories (or the semantic means of articulating category differences), rather than giving us plausible reason to think that these entities really are in the same

[32] E.g. Parfit, *On What Matters Vol 2*, p. 268.

[33] Here I shall understand a *category* as an ultimate class: the highest genera of entities in the world. Categories may contain species, but are not themselves species of any higher genera. If a set of categories is complete, then each entity in the world will belong to one and only one. Moreover, an attribute that can belong to entities in one category cannot be an attribute of entities in any other category. Whilst we understand that it would be false to say that the number 7 is even (for example), we would understand it as a *category-mistake* to say that the number 7 is red. We need not worry about having an exhaustive list of ontological categories in order to make use of the idea of category differences. See: F. Sommers, 'Types and Ontology', *Philosophical Review*, (1963); P.F. Strawson, 'Categories', in *Ryle: A Collection of Critical Essays*, ed. by O.P. Wood and G. Pitcher (London: Macmillan, 1970).

[34] Joshua Hoffman and Gary S. Rosenkrantz, *Substance among Other Categories* (Cambridge: Cambridge University Press, 1994).

[35] Reinhardt Grossmann, *The Categorial Structure of the World* (Bloomington, Indiana: Indiana University Press, 1983).

category. Perhaps this is so, but the burden would surely rest with those making this claim.

Parfit might also say that even granting that both natural facts and normative facts would, as true propositions, appear in the same *basic* category of *proposition* (or some cognate of this), they would then be located in distinct and non-overlapping *sub*-categories e.g. natural facts in ‘propositions that describe certain natural states of affairs’, and normative facts in ‘propositions that tell us what is right or what we ought or have reason to do’ (or similar). However, to make this claim would be to beg the question against the naturalist, who would maintain that the latter is actually a sub-category *within* the former.

This contrasts sharply with the case of heat and medieval kings, where, on at least two common systems, we find that *property* type entities (such as heat) appear in different, non-overlapping categories to *object* type entities (such as medieval kings). I can identify none on which property and object entities appear in the same category. Specifically, on E.J. Lowe’s system, *Objects* (subdivided into *Substances* and *Non-Substances*) constitutes one category, whereas *Properties* is distinct (a subcategory of both the *Modes* and *Attributes* categories).[36] And on Hoffman and Rosenkrantz’s system, *Property* constitutes its own category (as a subcategory of *Abstract*); whilst *Material Object* is a distinct category (as a subcategory of *Substance*, which in turn is a subcategory of *Concrete*). I suggest we would reach the same conclusion if we were to take a Ryle/Husserl approach to distinguishing categories, on which we substitute one expression for another and look for absurdity.

Accordingly, I would submit that if Parfit’s claim is that to identify normative and natural facts is to make a fundamental category mistake (of a similar kind and

[36] E.J. Lowe, *The Four-Category Ontology: A Metaphysical Foundation for Natural Science* (Oxford: Clarendon Press, 2006).

magnitude to identifying heat with medieval kings), then this is an implausible claim. To the extent that this claim depends upon the JTD intuition, then it inherits the problematic nature of (moral) intuitions in general (as described in section 3.4). To the extent that it depends upon an analogy with e.g. heat and medieval kings being obviously or self-evidently in different basic categories, then the analogy seems to break down under critical scrutiny, with it being far from obvious or self-evident that normative and natural facts are in different basic categories. Finally, to the extent that a justification of the analogy depends upon the claim that there is a property that normative facts have but that natural ones cannot have (thereby explaining the category mistake), viz. *normativity,* then it merely begs the question against the naturalist. Accordingly, I submit that Parfit's Normativity Objection misses the mark.

Moving on to Paakkunainen's second variant of the JTD objection to normative naturalism, the non-naturalist might argue that moral facts have a special normative importance that would go missing in an account such as mine, where *X*'s being morally right for *P* is identified with *X*'s best serving *P*'s true strongest desire. As she says:

> An act's being morally right or wrong, in contrast [to normativity involving instrumental goodness], seems very important indeed. Naturalists should explain how *X*'s being welfare-maximizing differs from other facts involving instrumental goodness, such as the fact that a given pebble is useful for decorating gardens, so that the former is a distinctively normatively important fact while the latter isn't. Otherwise moral facts' special normative importance has gone missing.[37]

In response, I would argue that it is hard to conceive of anything that could be more important to us than what will best serve our true strongest desires. Certainly, it has an

[37] Paakkunainen, 'The "Just Too Different" Objection to Normative Naturalism', p. 6.

obvious normative importance to us that, for example, the fact that a given pebble is useful for decorating gardens does not have.

The non-naturalist might press this idea of 'special normative importance' further though, with the special normative importance or authority of moral facts generally cashed out in terms of normative reasons to be moral. For example, as FitzPatrick says, morality appears to 'essentially involve strong reasons to conform to [its] standards' (that is, to act rightly and not wrongly).[38] As Paakkunainen puts it:

> If we had only very weak reasons, or no reasons at all, to do what's morally right or to avoid what's morally wrong, then morality would seem to lack the kind of authority on us that it intuitively has.[39]

However, I would submit that Goal Theory *does* have the kind of authority that morality intuitively has, and *does* give us strong reasons to conform to its standards. As I pointed out in section 3.2 (when affirming Joyce's authority requirement), if *P* morally ought to Φ (in circumstances *C*) on Goal Theory, then this will be because doing Φ (in circumstances *C*) will best serve the strongest desire of a fully rational and sufficiently informed version of *P* (i.e. *P*'s true strongest desire). Now, the HTR* states that an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. So, on the conjunction of Goal Theory and the HTR*, *P* will necessarily have at least *pro tanto* normative reason to do what they morally ought to do. Moreover, combining this with Proportionalism (on which, when a reason is explained by a desire, its weight varies in proportion to the strength of that desire, and to how well the action promotes that desire), we find that *P* will in fact have *strongest* reason

[38] FitzPatrick, 'Skepticism About Naturalizing Normativity: In Defense of Ethical Nonnaturalism', p. 576.

[39] Paakkunainen, 'The "Just Too Different" Objection to Normative Naturalism', pp. 6-7.

to do *Φ* in C. Appealing now to Parfit’s platitudes, we may then infer that these reasons are then *decisive*, and that acting in this way is what *P* has *most* reason to do, and therefore *ought* to do. Therefore, it seems that Goal Theory does indeed satisfy Fitzpatrick’s requirement for morality to generate strong reasons to conform to its standards.

The non-naturalist might still be unsatisfied here, insisting further that the strong reasons to be moral must be *categorical* ones, applying to agents regardless of their contingent desires, thereby making morality’s authority on us *inescapable*.[40] In that case, even if doing Φ in *C* would *not* best serve *P*’s true strongest desire (or serve this desire *at all*, or even serve to *any* extent *any* of *P*’s desires), then it may still be *right* for *P* to do Φ in *C*. However, I argued in some depth in chapter 3 that the case for categorical reasons is not made; and, in the absence of this, our default position should be to deny their existence (with them then being metaphysically extravagant). Moreover, as I argued in section 3.2, my account *does* still respect a plausible version of inescapability, viz. inescapability* — on which, if *P* morally ought to Φ, then *P* morally ought to Φ regardless of whether Φing satisfies any of her *unenlightened* desires. Combined with the authority that I have just discussed, I would argue that my account adequately captures the putative inescapable authority of morality, such that if *P* morally ought to Φ, then *P* has a reason for Φing, and *P* morally ought to Φ regardless of whether Φing satisfies any of her unenlightened desires.

Accordingly, if the thought were that naturalists must explain the internal connection between rightness and strong *categorical* reasons, and that they are supposedly unable to do this satisfactorily, then I would deny this requirement. I do explain in naturalistic terms how Goal Theory’s account of rightness yields very

[40] E.g. FitzPatrick, 'Skepticism About Naturalizing Normativity: In Defense of Ethical Nonnaturalism', p. 562.

strong (in fact decisive) reasons to act rightly, regardless of anyone's *unenlightened* desires — which I think is all that is really required in order to answer the missing normativity objection. I do not explain how these reasons are then categorical ones, for the very good reason that I deny that they are. If anyone still insists that morality's inescapable authority cannot be captured without *categoricity*, then I submit that the onus is on him or her to explain why.

Other naturalists have taken a different approach to mine. For example, in his 'end-relational' view of reasons, Finlay attempts to capture in naturalistic terms categorical reasons to be moral.[41] On his view, what it is for *p* to be a reason for *A* to *Φ* is for *p* to be an explanation of why *A*'s Φ-ing would be good; where what it is for *A*'s Φ-ing to be good is for it to be the case that pr(e|A Φ's) > pr(e|A doesn't Φ), where *e* is an end made salient by conversational context.[42] However, as others have noted, Finlay's view faces a number of serious objections. For example, it may generate categorical reasons for grossly *immoral* acts (e.g. Hitler probably had categorical reasons on Finlay's account to order the Holocaust, since many conversational contexts in Nazi Germany made relevant ends salient); it may fail to strictly capture *categorical* reasons to be moral (since, on Finlay's account, these reasons still depend upon *someone's* desires, albeit not the agent's); and it may fail to capture morality's inescapable authority.[43] Thus, as Paakkunainen observes, many anti-naturalists would take this as sufficient reason to reject Finlay's view, on the basis that it fails to capture the relevant normative appearances.

Trying to develop an account that respects these appearances better, we find Smith, who is developing a naturalist 'constitutivist' account of reasons, on which

[41] S. Finlay, 'The Reasons That Matter', *Australasian Journal of Philosophy,* (2006), 1-20; S. Finlay, *Confusion of Tongues: A Theory of Normative Language* (Oxford: Oxford University Press, 2014).
[42] Finlay, *Confusion of Tongues: A Theory of Normative Language*, pp. 40-41.
[43] Paakkunainen, 'The "Just Too Different" Objection to Normative Naturalism', p. 7.

reasons for action are grounded in what is constitutive of ideal agency. This account aims to capture the appropriate range of reasons to be moral for each agent, regardless of the agent's, or anyone else's contingent desires.[44] If it were combined with a buck-passing view of moral properties — analysing them in terms of normative reasons (as I do) — then it might perhaps ensure that morality essentially involves strong reasons to act rightly, where these reasons hold regardless of anyone's contingent desires. However, Smith's constitutivism has won few adherents, being criticised for failing to adequately justify the normative significance of ideal agency, for begging the question in the argument that certain desires are constitutive of ideal agency, and for generating reasons that do not neatly align with commonsense morality.[45]

Another option available to naturalists is to let go of some of the appearances that anti-naturalists insist on (e.g. categorical reasons), but to develop naturalist views that capture enough of the appearances to be, overall, plausible. Schroeder defends his hypotheticalism in this way. On hypotheticalism, reasons do depend on agents' contingent desires, in the following way: what it is for *p* to be a reason for *A* to Φ is for *p* to help explain why *A*'s Φ-ing would promote the object of some contingent desire of *A*'s. Schroeder's view turns upon the idea that acting rightly always promotes some of one's desires to some extent, regardless of what those desires are. However, as I have already pointed out (in sections 2.6 and 3.1), Schroeder's account may, amongst other problems, generate an explosion of agent-neutral reasons — something that Schroeder acknowledges as a potentially catastrophic problem for his account, and one that he is unsure it can resist.

[44] Michael Smith, 'A Constitutivist Theory of Reasons: Its Promise and Parts', *Law, Ethics and Philosophy,* (2013), 9-30; Michael Smith, 'The Magic of Constitutivism', *American Philosophical Quarterly,* (2015), 187-200.

[45] E.g. M. Bukoski, 'A Critique of Smith's Constitutivism', *Ethics,* (2016), 116–46; Enoch, *Taking Morality Seriously*.

Unlike Finlay and Smith's views, mine does not affirm categorical reasons, since I find these implausible. As such (like Schroeder's hypotheticalism), it gives up on some of the appearances upon which non-naturalists insist. However, unlike Schroeder's account, mine does not generate an explosion of agent-neutral reasons. Rather, it generates *pro tanto* reasons for agents to act in ways that serve their enlightened desires, with these reasons being decisive ones to act 'rightly' — thereby capturing in naturalistic terms the internal connection between rightness and strong (instrumental) reasons. Accordingly, I submit that my account improves upon rival naturalistic ones, plausibly surviving the missing normativity objection.

As Paakkunainen observes, there is a further aspect of the missing normativity variant of the JTD objection. I have been looking at whether my naturalistic account can capture morality's special normative importance via a naturalist account of strong reasons to be moral. However, there is a concern that naturalist accounts of reasons miss the peculiar normative importance or authority of reasons themselves. As such, they will ultimately miss the peculiar normative importance or authority of any connected moral facts as well. Paakkunainen expresses it thus:

> For Finlay, reasons to Φ are, roughly, explanations why Φ-ing promotes an end made salient in conversational context, where the end is usually made salient by speaker preferences. For Schroeder, reasons to Φ are analyzed in terms of promotion of ends desired by the agent. The worry is that the ingredients in these naturalist accounts seem insufficient to insert genuine normative authority into the picture.[46]

Those who press this objection may grant that naturalists can capture reasons to be moral in a deflationary sense. However, they would insist that naturalists fail to

[46] Paakkunainen, 'The "Just Too Different" Objection to Normative Naturalism', p. 8.

capture reasons in a genuinely authoritative sense. As such, instead of capturing reasons' special normative importance, they deflate or eliminate it.

But how would we determine if naturalist accounts fail to capture reasons' own normative importance or authority? What precisely is it that is supposed to be captured, and what would it be to capture it? Unfortunately, there seems to be little agreement here amongst non-naturalists. One suggestion (denied by some) for a hallmark of genuine normative reasons is the internalist view that agents must be capable of being motivated by the normative reasons that apply to them. Here, remember my definition of a normative reason on the HTR*, viz. an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent. As such, on my account, if agent *A* has a normative reason to do *x* in circumstances *C*, then it will necessarily be the case that doing *x* in *C* will serve some of the desires of a fully rational and sufficiently informed version of *A* (call this *A**). In light of this, if *A* does have normative reason to do *x* in *C*, then is *A* capable of being *motivated* by this normative reason?

Let me consider *A** first. If *A** believes that she has a normative reason to do *x* in *C*, then, by inference, she will believe that doing *x* in *C* will serve some of her desires. In that case, *A** will then have pre-existing desires, and a means-end belief about what action will serve those desires. Thus, on the HTM, where motivation requires the presence of a belief and an appropriately related and independently intelligible desire, *A** would then be *motivated* by the normative reason to do *x* in *C*.

However, what would we find for agent *A*, who, as a real-world agent, is not fully rational and sufficiently informed? Well, if *A* now believes that she has a normative reason to do *x* in *C*, then, by inference (and assuming she understands what a normative reason is), she will believe that doing *x* in *C* would serve some of the desires of a fully rational and sufficiently informed version of herself (i.e. *A**). Now, if we call *A*'s desire set *D*, *A**'s desire set *D**, and one of *A**'s relevant desires *d**, is it necessarily, or only contingently, the case that *d** will be present in *D* (where *d** being present in *D* means that *A* would be motivated to some extent do *x*, since she would then have a desire, *d**, and a means-end belief that doing *x* will serve *d**)? I have already carried out a similar analysis in section 2.4, so I shall not rehearse it in detail here. In summary, I think we would find that whether non-ideal agent *A* will necessarily be motivated (to some extent) to do *x* in *C* will turn upon our preferred account of the belief-desire process. If desires can be changed as the conclusion of reasoning only if a desire is among the premises of the reasoning (as Sinhababu argues, and I am persuaded), then I think it is plausibly the case that *A* will necessarily be motivated by the normative reason to do *x* in *C*. Otherwise, she will be only contingently motivated. However, even in this latter case, I would suggest that *A* likely *will* still be motivated (to some extent) by the normative reason. I say this because I think it is plausible that, upon suitable deliberation, most people would generate a *pro tanto* realizer desire to do what they know a fully rational and sufficiently informed version of them would have a desire to do in those circumstances (based upon a general standing desire to act rationally and informedly). In that case, then, in conjunction with the aforementioned means-end belief, they would be motivated to some extent to do *x* in *C*.

As a result, if we were to take as a hallmark of genuine normative reasons the internalist view that agents must be capable of being motivated by the normative reasons that apply to them, then I think my account plausibly satisfies this condition.

Another suggestion for a hallmark of genuine normative reasons is that they are those that it would be irrational to ignore in deliberation. However, as pointed out by Paakkunainen, it is unclear what would count as irrationality here (is it just instrumental inefficacy or incoherence, or might some moral failings count as irrationality too?); it is unclear why reasons being irrational to ignore is more important than reasons being e.g. immoral to ignore, or impolite to ignore, or against the club rules to ignore; and it is unclear that the normative import of reasons requires them to be taken into account in any kind of deliberation whatsoever.[47]

Notice that my account does seem to meet this hallmark on at least one plausible and popular conception of irrationality. Specifically, if irrationality is understood in instrumental, goal-oriented terms, whereby it is irrational to knowingly fail to do one’s best (or at least what one appropriately thinks is adequate) to achieve one’s ends, then it would seem to be irrational to ignore in deliberation what would serve some of the desires of a fully rational and sufficiently informed version of oneself, since it is hard to imagine that enlightened desire satisfaction would not feature in one’s set of ‘ends’ (even if other things feature too). However, for fear of begging the question (since the objection was phrased in terms of a contrast between genuine normativity and instrumental normativity), I shall bracket this conception. Accordingly, given the overall lack of clarity in what would count as irrationality, I shall set aside that putative hallmark.

Thus, I would submit that the target normative facts and properties on my account plausibly bear at least the former proposed hallmark of normative importance, naturalistically capturing reasons’ own normative importance or authority. Advances in the JTD debate may generate additional candidate hallmarks of genuine normative importance against which my account might be assessed (including perhaps a suitably

[47] Paakkunainen, 'The “Just Too Different” Objection to Normative Naturalism', p. 9.

clarified version of the second hallmark), but for now I think that my account is adequate on this front.

So, where does this leave the missing normativity objection to my account? Non-naturalists argue that certain normative properties, such as being morally right or wrong, good or bad, being what one ought to do, and so on, have a special kind of normative importance, robustness, and authority that goes missing in attempts to naturalise them, and that normative reasons also have a peculiar normative importance or authority that naturalist accounts of reasons inevitably miss. However, I find this claim to be no more *prima facie* rationally persuasive than the claim that heat has a particular quality (of 'heatiness', perhaps) that goes missing in attempts to naturalise it. Heat *just is* molecular kinetic energy, even if something might intuitively seem to go missing in attempts to naturalise it by identifying the former with the latter. Likewise, I would argue, with my proposed identity between normativity and natural facts and properties.

FitzPatrick thinks that 'any view that implies that genuine normativity can arise from mere psychology…seems to just be turning the idea of normativity into something else, such as motivation or instrumentality.'[48] Yet, to me, this is no more convincing than saying 'any view that implies that genuine heat can arise from mere energy…seems to just be turning the idea of heat into something else, such as molecular motion.' I would argue that the common naturalist strategy of reducing normativity to psychological facts can work if we choose the right psychological facts and the correct theory of normative reasons. Moreover, I think that I *have* chosen the right psychological facts (i.e. agents' own enlightened desires), and adopted the *correct* theory of reasons (i.e. the HTR*). Having done this, I think that my naturalistic

[48] FitzPatrick, 'Skepticism About Naturalizing Normativity: In Defense of Ethical Nonnaturalism', p. 584.

account of normativity then gives the non-naturalists all that they could reasonably ask for from the naturalist, in adequately capturing the desired robustness and authority of normativity by generating strong (in fact, decisive) reasons for agents to be moral, irrespective of their present, unenlightened desires. Yes, it does not generate *categorical* reasons to be moral — something many non-naturalists insist upon — but I deny that there are such reasons, making that requirement an unsatisfiable one. Within the framework of my discussion in section 3.1, I might say that a theory of normative reasons that incorporates categorical reasons then *overgenerates* reasons — counting some things as reason-giving that clearly are not.

As such, my account delivers the desired *authority* of moral reasons, but not their putative categorical *inescapability*. Yet, as I showed in section 3.2, my account does still deliver inescapability*. Moreover, it also gets us what Joyce posits as the *reason* why we actually want inescapability, viz. without inescapability, it would not necessarily be the case that agents who possess or acquire knavish desires ought to do what they morally ought to do. Therefore, I submit that my account captures enough of the appearances that the non-naturalist insists upon to be plausible overall, getting all the robustness and authority that is really required (and possible), without postulating queer and metaphysically extravagant entities. Thus, I think that I succeed in capturing the ‘normativity of the normative’ (contra to what Enoch supposes is possible for naturalist accounts).[49]

[49] Enoch, *Taking Morality Seriously*, pp. 104-08.

4.6 Conclusions

The aim of this chapter was to respond to the JTD objection to normative naturalism, according to which normative facts and properties are ‘just too different’ from natural facts and properties to be reducible or identical to them. If the objection were to succeed, then Goal Theory, with its naturalist account of normativity, would be undermined. This I have done, arguing that Goal Theory successfully resists it.

First, I introduced what some philosophers take to be the underlying motivation for non-naturalism, viz. the ‘just too different’ intuition, and explained why I think the bare intuition carries little weight. Next, I presented a positive argument for my particular variant of naturalistic normativity, arguing that all normative concepts are analysable in terms of one, fundamental normative concept, viz. a *reason*, and that this putative fundamental normative concept picks out a natural property. After that, I outlined the JTD objection to normative naturalism, and sharpened it into two main variants that may be targeted at Goal Theory’s account: one based upon missing *structure*, and the other based upon missing *normativity*. The former I cashed out in terms of the ‘argument from because’, maintaining that my account resists it, since my account’s reductive use of ‘because’ means that the *same* property is being referred to in two different ways. As such, the identity that goal Theory proposes becomes relevantly analogous to the identity between heat and molecular kinetic energy, and between water and H_2O — neither of which is vulnerable to (or ever targeted by) the argument from because.

In the second case, I first considered Parfit’s argument that normative and natural facts are in two quite different, non-overlapping categories — finding it to be unpersuasive. I then evaluated the objection that the target normative facts and properties have a special normative importance that somehow goes missing in attempts to naturalise them. In response to that, I argued that my account succeeds in capturing this special normative importance, generating strong (in fact *decisive*) reasons to act rightly and not wrongly that apply to agents regardless of their present, unenlightened desires. As such, I argued that it is suitably robust and authoritative. Moreover, I argued that my naturalist account of reasons captures the peculiar normative importance or authority of reasons themselves. Accordingly, I would argue that Goal Theory plausibly resists the JTD objection to normative naturalism.

Once again, whilst giving up on categoricity (justifiably, in my view), my account still captures enough of the appearances that are insisted upon by non-naturalists to be, overall, plausible. Specifically, it captures the special kind of normative importance, robustness, and authority that they want, and which they think goes missing in attempts to naturalise them. In addition, it satisfies the anti-naturalists' proposed hallmarks of genuine normative reasons, thereby plausibly accounting for the peculiar normative importance or authority of reasons themselves.

In so doing, I suggest that my account improves both upon other naturalist accounts — which either mistakenly (in my view) affirm categorical normative reasons, or else seem to suffer from an explosion of agent-neutral reasons — and upon non-naturalist accounts, which I suggest are undermined by the kinds of objections that I outlined in chapter 1.

Having now answered the second of the three dominant challenges levelled at accounts of Goal Theory's type, in the following chapter I shall critically evaluate the third, viz. that ethical egoism succumbs to a number of internal and external criticisms, spelling serious trouble for Goal Theory's egoist account.

Chapter 5

The challenge of ethical egoism

Do we have natural moral duties to other people simply because our actions could help or harm them? Should the interests of other people count in our moral deliberations? According to the commonsense view, the answer to both questions is yes. If there is one thing it feels we can say with certainty about morality, it is that it cannot be all about self-interest. In fact, morality would seem to be the very antithesis of self-interest, with anyone who claims otherwise appearing conceptually confused. Yet, once again, what if the commonsense view is mistaken?

The ethical egoist would challenge this commonsense view, claiming instead that moral agents ought to do what serves their own self-interest (I shall attenuate this conception later).[1] By this, they do not mean that agents should necessarily *avoid* actions that serve the self-interest of others. After all, there may be occasions where serving one's own self-interest permits or demands that one also serves the self-interest of others. However, in such cases, the egoist would say that what makes the act right is not any benefit that accrues to others, but the benefit that accrues to oneself. As with utilitarianism, egoism is a consequentialist theory, but where

[1] Ethical egoism is distinct from *psychological* egoism, according to which people do always act so as to best serve their own self-interests. The former is a normative theory about how we *ought* to act, whereas the latter is a descriptive theory about how people actually *do* act. Ethical egoism should also be distinguished from *rational* egoism, according to which an action is *rational* just in case it serves one's self-interest. I deny psychological egoism, but endorse a form of rational egoism. Hereafter, unless specified otherwise, whenever I refer to egoism I shall mean *ethical* egoism. The particular ethical egoist view to which I refer is a *universal* one, according to which everyone ought to serve his or her *own* self-interest. This is as opposed to the view of the *individual* egoist, who would claim that 'everyone ought to serve *my* self-interest.' The latter is not capable of being universalised, and nor would the individual egoist be likely to persuade others to adopt a moral system that benefits only the individual egoist herself. As such, I think it is probably an inadequate ethical theory.

utilitarianism is an *impartialist* ethical theory, egoism is a *partialist* one. Both theories require the maximisation of some maximand, but where consequentialism requires the maximisation of *total* utility, egoism requires the maximisation of one's *own* utility.[2]

The egoist view has some clear advantages. For example, it avoids any possible conflict between self-interest and morality, since to act morally is always to act in one's self-interest. Moreover, because, on egoism, morality always best serves one's self-interest, then egoists have a ready answer to the question of why they should be moral. Additionally, on the assumption that it is rational to pursue one's own interests, then it makes moral behaviour rational by definition. I have already shown (in sections 2.4 and 2.6, respectively) that, on Goal Theory in particular, anyone who sincerely holds a moral view will plausibly be motivated to some extent to comply with it, and will also have excellent reasons for so doing (at least on a Humean theory of reasons). By contrast, non-egoist views — on which acting morally need not serve agents' self-interest — struggle to supply us with reasons to comply, and may leave us with insufficient motivation to actually do so.

Notwithstanding the aforementioned benefits, egoism is an unpopular doctrine amongst philosophers. As Keith Burgess-Jackson observes, it is said to be 'refuted' by the kinds of arguments that would apply equally well to utilitarianism, and is taken far less seriously, and engaged with far less charitably, than that theory.[3] Moreover, egoism, unlike utilitarianism, is very often misrepresented, disparaged, and dismissed out of hand, with some philosophers being openly contemptuous and dismissive of it.[4]

[2] The similarities are discussed here: H. Sidgwick, *The Methods of Ethics*, 7th edn (Indianapolis, IL: Hackett Publishing Company, 1981 (1907)), p. 84.

[3] Burgess-Jackson, 'Taking Egoism Seriously'.

[4] For example, James Rachels has described it as 'a wicked view': James Rachels, 'Two Arguments against Ethical Egoism', *Philosophia,* (1974), 297-314 (p. 298). Whilst Holmes Rolston says that 'If moral philosphers [sic] have nearly agreed to anything, they agree that ethical egoism (I ought *always* do what is in my enlightened self-interest) is both incoherent and immoral': H. Rolston, *Environmental Ethics: Duties to and Values in the Natural World* (Philadelphia, PA: Temple University Press, 1988), p. 294.

Intuitively, it feels wrong to many, with morality seeming to be fundamentally about acting altruistically, and not at all about serving one's own self-interest. However, with reference to my earlier discussion of the evolutionary origins of our moral intuition (section 3.4), I suggest this may be a failure to distinguish between proximate and ultimate aims. Specifically, I would argue that the *ultimate* aim of our intuitive moral sense is to dispose us towards the kinds of behaviours that tend to increase our differential reproductive success (or, at least those that did so in the case of our Pleistocene hunter-gatherer ancestors). As such, our intuitive morality is in one sense fundamentally self-interested. However, it so happens that this ultimate aim is often best served by having the *proximate* aim of promoting the interests of others (and if it were not, then we probably would not have evolved a disposition towards promoting these interests).

Egoism is open to a number of serious challenges, including that it is collectively self-defeating, cannot deal with conflicts of interest, is logically inconsistent, is unacceptably arbitrary, and implies moral propositions that are false or unacceptable. Yet engaging with it honestly and charitably helps to raise the level of argumentation and analysis, enabling the theory to be usefully developed. As an egoist theory, on which agents morally ought to do what will best serve their true strongest desires (and where it is never the moral action to not best serve one's true strongest desire), Goal Theory faces the aforementioned challenges. Is it resistant to them, or instead fatally undermined? In what follows, I shall argue for the former. In so doing, I shall submit that at least one particular variant of egoism can be successfully defended, accruing all of the previously identified benefits, whilst succumbing to none of the criticisms. In that case, what might at first be seen as a weakness of Goal Theory (i.e. that it is an egoist theory), becomes one of its strengths. Accordingly, I aim to answer directly the challenge identified in chapter 1, viz. that ethical egoism succumbs to a number of internal and external criticisms, spelling serious trouble for Goal Theory's egoist account.

5.1 The case against ethical egoism

In order to present and respond to the case against ethical egoism, I shall follow a basic template outlined by Keith Burgess-Jackson. According to him:

> A theory, whether positive or normative, can be criticized in either (or both) of two ways. An *internal* criticism seeks to show that the theory is incoherent. A theory can be incoherent either because it has inconsistent implications (i.e., gives contradictory or contrary results, makes contradictory or contrary predictions) or because its components are not mutually supportive… An *external* criticism, by contrast, seeks to show that the theory has false or unacceptable implications. Just as anything that implies a falsehood is false, anything that implies an unacceptable proposition is unacceptable.[5]

In line with this, I shall first critically evaluate two familiar internal criticisms of ethical egoism. In response, I shall argue that neither of them successfully demonstrates that Goal Theory's particular variant of egoism is incoherent. Secondly, I shall address external criticisms. The general form of these criticisms may be represented by the following modus tollens argument:

1) Theory *T* implies proposition *p*.

2) *p* is false.

[5] Burgess-Jackson, 'Taking Egoism Seriously', p. 533.

C) Therefore, *T* is false.[6]

As Burgess-Jackson notes, we might take *T* to be utilitarianism, for example, and *p* to be the proposition that it is sometimes morally permissible to punish the innocent. Then, if one admits premise 1 with this particular *T* and *p*, but denies *p*, then one is logically committed to rejecting utilitarianism.

Of course, I am interested here in the case where *T* = Goal Theory. So, in the section on external criticisms I shall consider a representative case that can be framed in terms of some putative *p* that is supposedly being implied by Goal Theory, but where *p* is apparently false (thereby entailing the falseness of Goal Theory). In response, I shall challenge both premises, arguing that it is very improbable in the real world that my account, when properly understood and applied, actually implies the *p* in question; but that, if ever it did, I would then deny that *p* would be false *in those circumstances*. By such means, I shall argue that Goal Theory survives the challenges aimed at ethical egoism.

Before I proceed, I must first unpack and sharpen up my understanding of egoism. As a first pass, based upon my earlier statement, I would suggest the following conception:

> **Egoism$_1$**: the view that agents morally ought to do what serves their own self-interest.

However, this conception raises an obvious objection, viz. that there might be multiple, conflicting actions that would to some extent serve an agent's own self-

[6] I exclude Burgess-Jackson's 'unacceptability' condition, on the basis that it is unclear to me in what circumstances and to what extent a proposition — and thence the theory that generated it — being judged by some to be 'unacceptable' should diminish its epistemic status.

interest. Thus, on this conception, it may be the case that an agent morally ought to do *x* in circumstances *C*, and that they morally ought to do ~*x* in *C*. However, unlike with *reasons*, where I think it is perfectly coherent to have a reason to do *x* and a reason to do ~*x*, when it comes to what an agent morally ought to do I think that any coherent theory should not yield conflicting actions in any particular circumstances. In order to defuse this objection, I shall attenuate my conception thus:

> **Egoism$_2$**: the view that agents morally ought to do what *best* serves their own self-interest.

However, this conception still seems unsatisfactory, because it is unclear what is meant by 'self-interest' here. Accordingly, I shall endorse the widely held *desire* account, on which an agent's self-interest is identified with the satisfaction of their desires. (This contrasts with *objective* accounts, which identify self-interest with the possession of states, such as virtue or knowledge, which are valued independently of whether they are desired.) Thus, I shall further refine the above conception:

> **Egoism$_3$**: the view that agents morally ought to do what best serves their desires.

Yet this leads once again to a form of the earlier objection, insofar as an agent might have a desire that would be best served by doing *x* in circumstances *C*, and another desire that would be best served by doing ~*x* in *C*. In this case, I shall disarm the objection by focussing upon the agent's *strongest* desire:

> **Egoism$_4$**: the view that agents morally ought to do what best serves their strongest desire.

However, a couple of disambiguations are now called for. First, when I refer to an agent’s strongest desire, do I mean their strongest *present* desire or their strongest *enlightened* desire (where I understand the latter to be the desire of a fully rational and sufficiently informed version of the agent)? Second, is the desire-satisfaction in question at a particular time, or over the long term (perhaps over the agent’s whole life)? Of the possible combinations, I would like to highlight two:

> **Egoism$_5$**: the view that agents morally ought to do what best serves their strongest present desire at a particular time.

> **Egoism$_6$**: the view that agents morally ought to do what best serves their strongest enlightened desire over the long term.

Expressed in terms of self-interest, we have:

> **Egoism***: the view that agents morally ought to do what best serves their own self-interest* (where doing *F* is in an agent’s self-interest* if it will result in their strongest present desire being satisfied at a particular time).

> **Egoism****: the view that agents morally ought to do what best serves their own self-interest** (where doing *F* is in an agent’s self-interest** if it will result in their strongest enlightened desire being satisfied over the long term).

As will become clear, I think that if they bite, the standard objections to egoism only do so on something like egoism* (which more closely aligns with the common view that egoism is a fundamentally *selfish* doctrine). However, Goal Theory is equivalent to egoism**, on which egoism is a kind of enlightened self-interest, and where promoting the interests of others effectively becomes a subgoal of promoting one's own interests (as I shall show).[7] Thus, in what follows, I defend egoism** specifically, rather than any other variant of egoism.

5.2 Egoism cannot deal with conflicts of interest

The first of the internal criticisms of ethical egoism I shall consider comes from Kurt Baier:

> Let B and K be candidates for the presidency of a certain country and let it be granted that it is in the interest of either to be elected, but that only one can succeed. It would then be in the interest of B but against the interest of K if B were elected, and vice versa, and therefore in the interest of B but against the interest of K if K were liquidated, and vice versa. But from this it would follow that B ought to liquidate K, that it is wrong for B not to do so, that B has not 'done his duty' until he has liquidated K; and vice versa. Similarly K, knowing that his own liquidation is in the interest of B and therefore, anticipating B's attempts to secure it, ought to take steps to foil B's endeavours. It would be wrong for him not to do so. He would 'not have done his duty' until he had made sure of stopping B. It follows that if K prevents B from liquidating him, his act must be said to be both wrong and not wrong-wrong because it is the prevention of what B ought to do, his duty, and wrong for B not to do it; not wrong because it is what K ought to do, his duty, and wrong for K not to do it. But

[7] When discussing Goal Theory I generally omit any reference to agents' true strongest desire being best served *over the long term*. However, this should be assumed. For example, some action that might best serve an agent's true strongest desire only in the short-term, but frustrate it over the longer term (e.g. stealing some money), would not generally be commanded on Goal Theory.

> one and the same act (logically) cannot be both morally wrong and not morally wrong…
>
> This is obviously absurd. For morality is designed to apply in just such cases, namely, those where interests conflict. But if the point of view of morality were that of self-interest, then there could never be moral solutions of conflicts of interest.[8]

We may discern two distinct arguments from this passage. First, Baier argues that ethical egoism is self-contradictory — a very serious charge. According to him, it is *B*'s duty is to liquidate *K*, and *K*'s duty is to prevent *B* from doing it. However, it is wrong to prevent someone from doing his duty, and so it is wrong for *K* to prevent *B* from liquidating him. Thus, according to Baier it is both wrong and not wrong on ethical egoism for *K* to prevent *B* from liquidating him — hence the self-contradiction. However, as James Rachels points out, this argument is straightforwardly undermined by noting that ethical egoism is not committed to the proposition that it is wrong to prevent someone from doing his duty.[9] The ethical egoist would only endorse a qualified version of this proposition, whereby it is wrong for one to prevent someone from doing his duty *just in case* him doing his duty is in one's own best interests. Accordingly, I shall set this argument aside.

Second, Baier argues that ethical egoism cannot provide solutions to conflicts of interest, and so must be wrong, because providing such solutions is something that any adequate theory of morality must be able to do. Baier's argument may be reconstructed as follows (replacing egoism and self-interest in general with egoism** and self-interest** in particular, on the basis that it is the latter concepts that I am defending here):

[8] Kurt Baier, *The Moral Point of View* (Ithaca, NY: Cornell University Press, 1958), pp. 189-90.

[9] Rachels, 'Ethical Egoism', p. 198.

Argument 10

P1)	For a theory of morality to be adequate, it must be able to provide harmonious solutions for conflicts of interest.
P2)	B and K have a conflict of interest, insofar as it is in the self-interest** of B but against the self-interest** of K if B is elected president (and vice versa).
P3)	If it is in the self-interest** of B but against the self-interest** of K if B is elected president (and vice versa), then it is in the self-interest** of B but against the self-interest** of K if K were liquidated (and vice versa).
P4)	If it is in the self-interest** of B but against the self-interest** of K if K were liquidated (and vice versa), then it follows on egoism** that B ought to liquidate K, that it is wrong for B not to do so (and vice versa).
C1)	Therefore, on egoism**, B ought to liquidate K, and it is wrong for B not to do so (and vice versa). [P2, P3, P4]
P5)	But K, knowing that his own liquidation is in the self-interest** of B, ought on egoism** to take steps to foil B's endeavours, with it being wrong of him not to do so.
C2)	Therefore, on egoism**, B ought to liquidate K, and K ought to stop him (and vice versa). [C1, P5]
P6)	If, on egoism**, B ought to liquidate K, and K ought to stop him (and vice versa), then egoism** does not provide a harmonious solution for B and K's conflict of interest.

C3)	Therefore, egoism** does not provide a harmonious solution for B and K’s conflict of interest. [C2, P6]
C4)	Therefore, egoism** is not an adequate theory of morality. [P1, C3]

Here I shall understand a *conflict of interest* to be a situation in which the self-interests** of two (or more) agents are incompatible, such that what is in the self-interest** of one goes against what is in the self-interest** of the other(s). So, for example, if it is in the self-interest** of *B* but against the self-interest** of *K* if *B* is elected president (and vice versa), then, on my understanding, *B* and *K* will have a conflict of interest. Accordingly, I am prepared to grant premise P2 of Argument 10. I shall take a ‘solution’ to the conflict of interest to be a set of actions that resolves the conflict of interest, such that the agents’ self-interests** are no longer incompatible. Note that Baier talks of a ‘moral’ solution to conflicts of interest (which could be interpreted as meaning nothing more than a solution provided by our chosen moral theory). However, in line with Rachels (and as I find it more useful), I shall understand this to mean a *harmonious* solution. This is reflected in my reconstruction of Argument 10.

With regard to the conflict of interest in question, observe that we may model the interaction between *B* and *K* as a prisoner’s dilemma game, within which each player is pursuing their own self-interest**, and this self-interest** is affected not only by what they do but also by what the other player does.[10] Since the prisoner’s dilemma is so familiar, I can afford to be quick. In game theory parlance, participants in a

[10] For more on the prisoner’s dilemma and its applications, see, for example: Robert Axelrod, 'The Emergence of Cooperation among Egoists', *The American Political Science Review*, (1981), 306-18; Axelrod, *The Evolution of Cooperation*; Kenneth Binmore, *Playing Fair: Game Theory and the Social Contract*, Vol. 1 (Cambridge, MA: MIT Press, 1994).

prisoner's dilemma either cooperate or defect. In this case, defecting would be constituted by one player attempting to liquidate the other (and attempting to foil the other player's attempts to liquidate him) in order to take the presidency. By contrast, cooperating would be understood as being some kind of cooperative strategy, such as *B* and *K* agreeing to hold and abide by the result of a fair presidential election.

For simplicity, I shall assume for now that if both of them defect, then neither *B* nor *K* would be more likely to succeed in his attempts to liquidate the other, and that they will both continue in their attempts until one succeeds. As such, if both *B* and *K* defect, then we should assign a 50% probability to *B* liquidating *K*, and a 50% probability to *K* liquidating *B*. Therefore, from *B*'s perspective, defecting when *K* also defects carries a 50% chance of him getting what he wants (i.e. the presidency) and a 50% chance of being liquidated. I shall ignore any costs associated with losing in a (cooperate, cooperate) scenario (e.g. financial ones), since these are likely to be insignificant compared to the cost of being liquidated. Obviously, cooperation in the sense described (beyond not attempting to liquidate one's opponent) assumes that communication between the players is possible. Any agreements are assumed to be nonbinding though, to allow for the possibility that players may defect by reneging on their agreements. We now have four possible outcomes:

> **Option 1**: *B* cooperates whilst *K* defects. On a simple two-player one-off game, this is the best possible outcome for *K* (i.e. guaranteeing him the presidency), and the worst possible outcome for *B* (i.e. certain death). From *B*'s perspective, this option would constitute the 'sucker's payoff'.

> **Option 2**: *B* defects whilst *K* cooperates. On a simple two-player one-off game, this is the best possible outcome for *B* (i.e. guaranteeing him the presidency), and the worst possible outcome for *K* (i.e. certain death). From *B*'s perspective, this would constitute the 'free-rider' position.
>
> **Option 3**: Both *B* and *K* defect. On a simple two-player one-off game, this is the third-best outcome for both *B* and *K*, with each having a 50% chance of winning the presidency and a 50% chance of being liquidated.
>
> **Option 4**: Both *B* and *K* cooperate. On a simple two-player one-off game, this is the second-best outcome for both *B* and *K*. Each has a 50% chance of securing the presidency, with no significant costs associated with losing. This option is then the *Pareto-optimal* outcome, insofar as no one can become better off without someone becoming worse off. (More precisely, there is no other outcome that is strictly preferred by at least one player that is at least as good for the other.)[11]

Now, if *B* and *K* are both egoists**, then what ought they to do? Consider things from *B*'s perspective. Either *K* will cooperate or *K* will defect. If *K* cooperates, then the best option for *B* would appear to be defection (giving *B* the best possible outcome). However, if *K* defects, then the best option for *B* is still to defect (avoiding certain death, and giving him a 50% chance of getting the presidency). Thus, whatever *K* does, defection would seem the best option for *B*. Consequently, defection is a

[11] On Pareto optimality, see, for example: Allen Buchanan, *Ethics, Efficiency, and the Market* (Totowa, NJ: Rowman & Allanheld, 1985).

dominant strategy for *B*, insofar as it would appear to be the best strategy for him to adopt, regardless of what *K* does.

Yet *K* will also be going through the same process of reasoning, reaching the same conclusion. As a result, *K* will also defect. But when each party chooses his dominant strategy (i.e. defection), then an equilibrium is produced that is the third-best result for both (and Pareto-suboptimal, insofar as there is another outcome that is strictly preferred by at least one player that is at least as good for the other). By contrast, if *B* and *K* had both cooperated, then they would have produced the Pareto-optimal equilibrium. This is then the dilemma: if each player adopts what seems to be his best choice as an egoist**, then both players do worse than if they had collectively acted benevolently by cooperating.

Framed as a prisoner's dilemma game, there are now three in principle possibilities that might obtain, contingent upon the particular circumstances in which the players find themselves:

1. Cooperation will best serve the self-interest** of both players.
2. One player's self-interest** will be best served by defection.
3. Defection will best serve the self-interest** of both players.

I shall argue that, in all three cases, egoism** provides a solution to the conflict of interest. In the first case (which I suggest will obtain in almost all real-world cases), egoism** would dictate that both players do indeed cooperate, meaning that their self-interests** would then be compatible, and the conflict of interest would thereby be resolved. For the second (much rarer) case, I shall argue that withdrawing from the contest would best serve the self-interest** of the non-defecting player, and so this is what egoism** would dictate. Accordingly, there will no longer be an incompatibility between *B* and *K*'s self-interests**, and the conflict of interest is once again resolved. These two solutions are immune to Baier's argument, because in neither case does egoism** demand that the players ought to defect (with each having a moral duty to liquidate the other). As such, these solutions to the conflict of interest are both *harmonious* ones.

The last case (which I shall grant is in principle possible) is different, insofar as egoism** demands here that both players do in fact defect, with each having a moral duty to liquidate the other. Accordingly, this is not a harmonious solution, and so falls foul of premise P6 of Argument 10. However, in this case I shall deny premise P1. If I am right, then this is still a solution of sorts, and would still be immune to Baier's argument.

In conclusion, I shall argue that, for each possible variant, egoism** provides an adequate solution to the conflict of interest in question, and therefore Baier's argument fails to defeat egoism**. I shall now examine each variant in turn.

5.2.1 The cooperative solution

With defection being the dominant strategy for egoists** *B* and *K* in a two-player one-off game like the one described, the rational strategy for each player would appear to be to defect. (In line with the standard practice in game theory, as well as in economic theory, I assume that the rational choice for a player from amongst a set of possibilities is the one that *maximises* their expected utility — a concept that I shall discuss shortly.) So, is Baier vindicated?

Observe that if there was a sufficient penalty attached to defecting, then the dominant strategy for each player might change. Specifically, in such circumstances, it might become the case that, whatever *K* does, *B* is better off cooperating (and likewise for *K*). In that case, the Pareto-optimal outcome would be collectively motivated. It is my contention that in the real world, as opposed to the idealised two-player one-off game considered, this is what we would find. Here, *B* and *K*'s game is more accurately modelled as an n-player iterative one, with *B* and/or *K* continuing to interact with other players (and perhaps each other) after the interaction described. Moreover, in the real world, there are emotional, social, and legal penalties attached to defection (in addition to payoffs associated with cooperation). Once the game is modelled in this way, and the aforementioned penalties (and payoffs) taken into account, then I would suggest that cooperation becomes the optimal strategy for both players from a purely self-interested** perspective.

In terms of the aforementioned emotional penalties and payoffs, from *B*'s perspective (and likewise for *K*), with the knowledge that he has liquidated *K*, he would likely suffer from guilt, shame, remorse, and so on, even where it is in his interest that *K* is liquidated (per premise P3). Conversely, by cooperating, *B* may accrue an emotional reward, such as the joy and fulfilment of compassion and compersion.

With respect to the social penalties and payoffs, in an n-player iterative game, other players might become aware of how *B* and *K* have acted, remember this, and react accordingly. As Robert Axelrod demonstrated, in such a game, where players can remember their opponent's previous actions and alter their strategy accordingly, attenuated cooperation becomes the optimal strategy from a purely self-interested perspective (e.g. initial cooperation, but attenuated by then following 'tit-for-tat' behaviour, with some degree of forgiveness).[12] In this case, if other players became aware that *B* had defected (especially by liquidating his opponent), then he would acquire a bad reputation, and they would be disinclined to cooperate with him in the future. This lack of cooperation might take a number of forms, including shunning and retaliation. If *B* is shunned, then he forgoes (to some extent) the benefits of future indirect reciprocity. In terms of retaliation, others might try to avenge *K*'s death by killing *B*, or *B* might fall victim to a deadly coup. By contrast, if he cooperates with *K*, then *B* avoids retaliation, and fosters a good reputation, enabling him to reap the future rewards of direct and indirect reciprocity.

Finally, with respect to the legal penalties, if his offense is discovered, then *B* will likely face serious criminal charges for liquidating *K*, which might result in life imprisonment (or even execution). I think that in stating that it is in *B*'s interests to win the presidency even by liquidating *K*, Baier has simply failed to take into account a host of real-world factors. A goal can be in one's interests without any means of achieving that goal being in one's interests.

To give a better sense of the nature and magnitude of the optimal and sub-optimal strategies, let me employ rational choice theory in order to calculate an expected utility for each option. On this theory, we define expected utility thus:

$$EU(A) = \sum PA(o) \times U(o)$$

Here, $EU(A)$ is the expected utility of some act A; $PA(o)$ is the probability of outcome o conditional on A; and $U(o)$ is the utility of o.[13]

[12] Axelrod, *The Evolution of Cooperation*.

[13] For more on rational choice theory and utility maximisation, see: Cristina Bicchieri, 'Rationality and Game Theory', in *The Handbook of Rationality* (Oxford: Oxford University Press, 2003).

In order to calculate the expected utility, I must first assign utilities to each of the possible outcomes. My utility assignments will reflect the degree to which I think an act is in the self-interest** of the player concerned. Here I shall make the plausible assumptions that while there is utility for *B* in getting the presidency, this utility is far less than the *disutility* for him of being liquidated. (I shall do the analysis from *B*'s perspective, but a similar result would obtain if it were done from *K*'s perspective instead.) Accordingly, I shall take the utility for *B* of getting the presidency to be 2, and the utility of being liquidated as -10. Moreover, for the reasons explained, I shall assume that defectors will each receive a 5 point punishment (constituted by the conjunction of the aforementioned emotional, social, and legal penalties), and that cooperators will receive a 1 point payoff (if they have not already been liquidated, obviously). Furthermore, I shall now assume that the cooperating player in a (cooperate, defect) game has a 50% chance of not being liquidated, on the basis that he may alert the police, resulting in him being saved from liquidation, and the defector being disqualified from the presidential contest. In light of this, then, from *B*'s perspective, the calculation would proceed as follows:

Option 1: *B* cooperates whilst *K* defects: EU = ∑ (utility x probability) = (-10 x 0.5) + ((1 + 2) x 0.5) = -3.5

Option 2: *B* defects whilst *K* cooperates: EU = (2 x 0.5) - 5 = -4

Option 3: Both *B* and *K* defect: EU = ((2 - 5) x 0.5) + (-10 x 0.5) = -6.5

Option 4: Both *B* and *K* cooperate: EU = (2 x 0.5) + (0 x 0.5) +1 = 2

Notice that option 4 (cooperate, cooperate) is now not only the Pareto-optimal outcome (by some margin), but also the *Nash equilibrium* (i.e. the only outcome from which each player could only do worse by unilaterally changing strategy).[14] Moreover, cooperation is now the *dominant* strategy for *B* (and likewise for *K*) — since, whatever *K* does, *B* is better off cooperating.

One might object here that the numbers have been 'fixed' in order to give the desired result. However, observe that we still get the same outcome (albeit by a reduced margin) if the penalty for defection is reduced from 5 down to 1, and the payoff for cooperation is halved.[15] If we were to eliminate any penalty at all for defection, as well as any benefit for cooperation, and sufficiently reduce the chances that a cooperating player in a (cooperate, defect) game has of not being liquidated, then we would indeed change the outcome. However, those (rare) circumstances (which might equate to egoists** finding themselves in a local or global state of anarchy) would then fold into one of the two sets of circumstances that I shall be considering next, where egoism** would dictate either that a cooperative player withdraws, or else that both players try to win at all costs.

It might also be objected that the real world is not populated by egoists**, but by agents who often act uninformedly and irrationally, and who tend to consider only their short-term self-interests — and so more closely approximate egoists* than egoists**. (An extreme example of this would be criminals, who would usually best serve their strongest long-term enlightened desires by cooperating, e.g. by obeying the

[14] John Nash, 'Equilibrium Points in N-Person Games', *Proceedings of the National Academy of Sciences*, 36 (1950), 48-49.

[15] In that case, we get: option 1: *B* cooperates whilst *K* defects: EU = utility x probability = (-10 x 0.5) + ((0.5 + 2) x 0.5) = -3.75; option 2: *B* defects whilst *K* cooperates: EU = (2 x 0.5) - 1 = 0; option 3: both *B* and *K* defect: EU = ((2 - 1) x 0.5) + (-10 x 0.5) = -4.5; option 4: both *B* and *K* cooperate: EU = (2 x 0.5) + (0 x 0.5) + 0.5 = 1.5.

law, but who routinely defect anyway.) And the result of this is that *B* cannot entirely trust that *K* will not defect, even though doing so would not best serve his self-interest** (and vice versa). Whilst that is true, and a concern about relatively small defections might indeed be warranted, I suggest that almost all people in those circumstances would be disinclined to defect by attempting to liquidate their opponent, being fearful of the punishment that would likely await them, and tending to unquestioningly conform to the social norms forbidding murder (I shall return to the subject of social norms shortly). Any exceptions to this would generally be known about, meaning that an egoist** could then choose to either withdraw or to fight to win at all costs, depending upon the circumstances (with both of these folding into the next two sections).

As a result, once we model *B* and *K*’s game in a way that more closely approximates reality, then we obtain the result that each player pursuing his own self-interest** exclusively implies on egoism** that they ought to cooperate. This result runs counter to a common view that egoists will always defect in prisoner’s dilemma situations, on the basis that this is the selfish act, and egoism is inherently selfish. Even if it is true that an egoist* will always (or at least routinely) defect in a game like this, I am defending an egoist** position here. On egoism**, promoting the interests of others effectively becomes a subgoal of promoting one’s own interests.

This goes some way to answering the criticism that the egoist’s position is a ‘self-effacing’ one, whereby it behoves the egoist to avoid promoting egoism in public and to keep her true ethical beliefs a secret. If the real world was populated by strict altruists, then it would indeed be in the self-interest** of any egoist** to avoid promoting egoism** in public and to keep her true ethical beliefs a secret. However, such strict altruists are a rarity, with egoists* (or similar) being far more numerous.

However, egoists** would do better in a world populated by egoists** than one populated by egoists*, so it follows that they ought generally to promote egoism**.

Accordingly, we have a solution to *B* and *K*'s conflict of interest. Contra Baier, I would argue that, with regard to the conflict of interest in question, it need not be the case the egoism** would dictate that each player ought to liquidate the other. Rather, I would argue that, in the real world, egoism** would generally dictate that *B* and *K* ought to cooperate with each other.

5.2.2 The withdrawal solution

In light of the kinds of reasons previously adduced for why cooperating will generally be in the best self-interest** of both *B* and *K*, under what circumstances might one player's self-interest** be best served by defection? Based upon the previous expected utility calculation, I think that any in principle case would demand that the player in question faces only a negligible punishment for defection, gains no real payoff from cooperation, and is almost certain to succeed in liquidating any opponent. Can we conceive of any real-world circumstances that meet these criteria? I shall return to this later (in relation to external criticisms of egoism), but if the player in question is some kind of despot — lacking empathy, being largely unaffected by the disinclination of others to cooperate with him, able to act above the law, and effectively guaranteed to liquidate any person who runs against them for the presidency — then it may in principle be in their best self-interest** to always defect. Revising the expected utility calculation accordingly, we now have (taking *B* as the despot):

Option 1: *B* cooperates whilst *K* defects: EU = ∑ (utility x probability) = (-10 x 0.5) + (2 x 0.5) = -4

Option 2: *B* defects whilst *K* cooperates: EU = (2 x 1) = 2

Option 3: Both *B* and *K* defect: EU = (2 x 1) + (-10 x 0) = 2

Option 4: Both *B* and *K* cooperate: EU = (2 x 0.5) + (0 x 0.5) = 1

Now we see that defection becomes *B*'s dominant strategy, being the rational thing for him to do regardless of what *K* does. (In section 5.4, I shall consider some plausible outweighing costs of being a despot, but will set those aside for this purpose.) However, what do we get if we turn the calculation around, viewing it from *K*'s perspective? Given that *B* will always defect, I shall list just those options:

Option 1: *K* cooperates whilst *B* defects: EU = utility x probability = (-10 x 1) + ((1 + 2) x 0) = -10

Option 3: Both *K* and *B* defect: EU = ((2 - 5) x 0) + (-10 x 1) = -10

These figures merely confirm what is already obvious, viz. whether *K* cooperates or defects, if he competes with *B* for the presidency then he faces certain death. Observe, however, that there is another option, viz.

Option 5: *K* withdraws: EU = (2 x 0) + (-10 x 0) = 0

This option now has a higher expected utility for *K* than continuing with the game and either cooperating or defecting. Thus, in the absence of some reliable means to motivate or enforce mutual cooperation, it is then in *K*'s best self-interest** to *withdraw*. Accordingly, this is what egoism** would dictate, meaning that we once again have a solution to *B* and *K*'s conflict of interest. (And remember that on my definitions of self-interest** and egoism**, this will also be what *K* *morally* ought to do, not merely what he *prudentially* ought to do.)

5.2.3 The win at all costs solution

I think that in the last two sections I have covered almost all conceivable real-world situations in which *B* and *K* might compete for the presidency. Within the context of Baier's thought experiment, I suggest that any real-world situations in which this last option would apply would be very rare — since, from the previous expected utility calculations, such a situation would necessitate no real punishment for defection and nothing to gain from cooperation (as before), in addition to no possibility of withdrawal (or withdrawal being hugely costly). I think one can conceive of possible (but very rare) scenarios that in principle might meet these conditions. For example, two people, who are indifferent to each other's welfare, becoming stranded in a remote location (e.g. on a life raft, or in the wilderness), having only enough resources for one to make it home alive, and each being able to kill the other without the crime being discovered when they make it back to civilisation. Therefore, taking it to be in principle possible, then, revising the expected utility calculation accordingly, we have (from *B*'s perspective):

Option 1: *B* cooperates whilst *K* defects: EU = ∑ (utility x prob.) = (-10 x 1) + (2 x 0) = -10

Option 2: *B* defects whilst *K* cooperates: EU = (2 x 1) + (-10 x 0) = 2

Option 3: Both *B* and *K* defect: EU = (2 x 0.5) + (-10 x 0.5) = -4

Option 4: Both *B* and *K* cooperate: EU = (2 x 0.5) + (0 x 0.5) = 1

Observe now that the dominant strategy for *B* is *defection*, since, whatever *K* does, *B* does better by defecting. However, each player choosing his dominant strategy will produce a Pareto-suboptimal equilibrium that is the third-best result for both parties. By contrast, if *B* and *K* cooperated, then they would produce the Pareto-optimal equilibrium. However, in the absence of any penalty or payoff to enforce or motivate this Pareto-optimal outcome, then the rational strategy for each player would appear to be to defect. Since *K* will defect, leading to a (defect, defect) outcome, then *B* would do better by withdrawing (producing an expected utility of 0, as opposed to -4). However, as indicated, I assume here that a player cannot withdraw (at least not without incurring some outweighing cost, e.g. the player concerned being killed). Hence, in this scenario, egoism** would demand that *B* ought to liquidate *K* (and vice versa). Once again, I would argue that this constitutes a solution (of sorts) to the conflict of interest, with the conflict being resolved when *B* or *K* is eventually victorious.

Now, per premise P6 of Argument 10, Baier would want to argue that this is not a *harmonious* solution to the conflict of interest; and so, if such a case can obtain, then egoism** would be vulnerable to his argument. However, I also deny premise P1, on the basis that I do not accept that a theory of morality must necessarily be able to provide *harmonious* solutions to all possible conflicts of interest for it to be deemed adequate. If that was a necessary condition for adequacy, then a moral theory would be deemed inadequate if it demanded of an innocent child being tortured and unable to flee or seek help that they ought to try to kill (or at least disable) their torturer if they have an opportunity to do so, on the basis that the conflict of interest between child and torturer would then not be resolved harmoniously. Yet I imagine that few, including Baier, would be willing to concede that. Perhaps if a moral theory routinely failed to provide harmonious solutions to conflicts of interest, then we might justifiably question its adequacy, but this does not apply in the case of Goal Theory, which, as an egoist** theory, will (I would argue) almost always produce a harmonious solution.

It might seem strange to call winning at all costs a 'solution' to the conflict. However, there is some precedent here, with both James Rachels and John Hospers also identifying this as a possible solution to *B* and *K*'s conflict of interest (in fact, this is Rachels' primary response to Baier's argument, on behalf of the egoist). As Rachels says of the egoist:

> For him, life is essentially a long series of conflicts in which each person is struggling to come out on top: and the principle he accepts – the principle of Ethical Egoism – simply urges each one to do his or her best to win.[16]

[16] Rachels, 'Ethical Egoism', p. 197.

And Hospers says of Baier's example that:

> [The impersonal egoist's] view is that he should pursue his own interest exclusively, that *B* should pursue *B*'s, that K should pursue *K*'s, and so on for everyone else. What will he say in the case of *B* and *K*? He will advise *K* to try to win out over *B* by whatever means he can, and will advise *B* to try to win out over *K* by whatever means he can: in other words, to settle the thing by force or craft, and may the strongest or cleverest man win… His view does not, of course, provide a rational means of settling the conflict of interest, but it does provide a means. [17]

I think that the above characterisations would be mistaken for egoism**, on which enlightened cooperation, rather than winning by any means, will generally be demanded. However, I allow that there may in principle be certain circumstances in which egoism** would mandate this latter strategy.

Moreover, if we deem this to not be a genuine 'solution' to the conflict of interest, thereby rendering egoism** an inadequate theory of morality on Baier's argument, then notice that the same problem confronts non-egoist theories too. In a possible circumstance in which *K* will always defect, and *B* cannot withdraw, then what action would some non-egoist theory *T* dictate for *B*? There are only two possibilities: cooperate and face certain death, or try to win at all costs. If the latter option is deemed not to be a genuine solution to *B* and *K*'s conflict of interest, then this leaves only the former option. However, any *T* that dictated this would surely be inadequate. One might argue that a moral theory should allow for the possibility of sacrificing one's life for the benefit of others (e.g. to protect one's child, or as an act of heroism), but this situation is relevantly different to that one, with *B* having no outweighing interest in helping *K*. And egoism** would actually permit a sacrifice of

[17] John Hospers, 'Baier and Medlin on Ethical Egoism', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 12 (1961), 9-16 (p. 13).

the aforementioned kind, if the costs for agent concerned of continuing to live outweighed those of dying. Therefore, I would argue that either *T* will always be inadequate, or we must grant that winning at all costs *does* constitute a genuine solution to *B* and *K*'s conflict of interest. If we accept the former, then a problem facing all theories cannot be used as a means to eliminate only one of them. And, if we accept the latter, then egoism** is left intact.

In conclusion, and contra the sub-argument from P2–C1 of Argument 10, I suggest it will rarely, if ever, be the case that, on egoism**, *B* ought to liquidate *K* (and vice versa). On neither of the first two identified solutions to *B* and *K*'s conflict of interest is this so, yet these collectively exhaust almost all real-world possibilities. It is only on the third solution that the sub-argument from P2-C3 might run through. However, even then, I think Baier's overall argument fails, since I also deny premise P1. Thus, I would argue that Baier has not made the case that egoism** is wrong because it cannot provide suitable solutions to the conflicts of interest that he identifies.

5.3 Egoism is collectively self-defeating

According to another argument, egoism is collectively self-defeating. Consider the following thought experiment. *X* prefers to drive to work, rather than taking the bus. However, if everyone drives, then the traffic will become very congested. Therefore, everyone is better off in a situation where everyone takes the bus, rather than in a situation where everyone drives. However, *X* reasons that her driving to work has an infinitesimal effect upon the overall volume of traffic; so, because she prefers to drive, then it is in her self-interest to continue doing so. Yet everyone else reasons in the same way, leading to a situation that is worse for all. Formally:

Argument 11

P1)	Commuter X can make a decision to drive to work or to take the bus.
P2)	X prefers to drive.
P3)	X driving into work will have an infinitesimal effect upon the overall volume of traffic.
C1)	Therefore, it is in X's self-interest to drive.
P4)	But the same reasoning applies to (most of) the other commuters.
C2)	Therefore, it is also in their self-interests to drive.
P5)	But if (almost) everyone drives, then it will lead to terrible traffic congestion (as well as pollution).
P6)	It is nobody's self-interest to have terrible traffic congestion (or pollution).
C3)	Therefore, each commuter acting in his or her own self-interest leads to an overall result that is in nobody's self-interest.
C4)	Therefore, egoism is collectively self-defeating.

This challenge for egoism is related to Garrett Hardin's so-called 'tragedy of the commons', which highlights the conflict between individual and collective rationality.[18] It can be understood as an n-player version of the prisoner's dilemma (where n > 2). Following the previous approach, let me once again assign values to the various outcomes. For simplicity, I shall analyse this as a three-player game — with players *X*, *Y*, and *Z* — and make the assumption that this is sufficient to generate the unwanted traffic congestion. (In practice, of course, *n* would be in the thousands, if not the millions, but this simplified version illustrates the principle.) If all players cooperate — restricting their car usage, by sometimes using the bus instead — then they all get 3 points. If they all defect — continuing with unrestricted car usage, and thereby generating the unwanted traffic congestion and pollution — then they all get 2 points. If two players cooperate, and one defects, then the cooperators get 2 points and the defector gets 4 points. And if one player cooperates, but two players defect, then the cooperator gets 1 point and the defectors get 3 points. Accordingly, from *X*'s perspective we now have:

Option 1: *X* cooperates whilst *Y* and *Z* defect: 1 point

Option 2: *X* and *Y* cooperate, whilst *Z* defects: 2 points

Option 3: *X* and *Z* cooperate, whilst *Y* defects: 2 points

Option 4: All players cooperate: 3 points

Option 5: *X* defects, whilst *Y* and *Z* cooperate: 4 points

[18] Garrett Hardin, 'The Tragedy of the Commons', *Science*, 162 (1968), 1243-48. In this thought experiment, ranchers may graze their animals on a common field. The rational thing for each rancher to do is to add more and more livestock, in order to increase profits. However, since all ranchers will reason in the same way, the field will become overconsumed, and so no rancher will be able to graze it.

Option 6: *X* and *Y* defect, whilst *Z* cooperates: 3 points

Option 7: *X* and *Z* defect, whilst *Y* cooperates: 3 points

Option 8: All players defect: 2 points

Once again, we find that defection is the dominant strategy for *X*, as the player is better off choosing this strategy no matter what *Y* and *Z* do. Precisely the same reasoning will apply from *Y* and *Z*'s perspectives. However, when all players choose their dominant strategies, then they produce an equilibrium that is the third-best result for all (hence the dilemma). The Pareto-optimal outcome would be the (cooperate, cooperate, cooperate) outcome, as that is the outcome for which there is no other outcome strictly preferred by at least one player that is at least as good for the others.

In my previous analysis of *B* and *K*'s rivalry for the presidency, I described several factors that may attach an additional, outweighing cost to defection (e.g. emotional, social, or legal ones). However, in this case, there is typically little stigma attached to driving, so those who drive will generally incur no significant loss of reciprocity from others. Moreover, few would suffer any significant negative emotional impact from driving, and there are generally no legal penalties to doing so. As a result, we observe in the real world that very many people choose to drive, creating terrible traffic congestion — an outcome that is worse for everyone. What can be done in a case like this?

If there was a way to collectively encourage or enforce the Pareto-optimal outcome, then this would produce an outcome that would better serve *X*, *Y*, and *Z*'s self-interest** (moving them from their third best outcome to their second). In being the Pareto-optimal outcome, any further improvement for one could only be had at the expense of making one or more of the others worse off. So, how might we achieve this? One solution might take the form of an external Leviathan (e.g. government) exercising central control over car usage, by accurately determining and assigning the optimum car usage limits, monitoring people's compliance with these limits, and sanctioning noncompliance.[19] Imagine that the players would consent to such a scheme (I explain why below), and assume for simplicity that the external Leviathan has reliable and valid information, and is able to correctly and effectively impose penalties for defection. In that case, we might say, for example, that all defectors will receive a 2-point punishment, and nobody who does not defect will receive this punishment. Now, from *X*'s perspective we have:

Option 1: *X* cooperates whilst *Y* and *Z* defect: 1 point

Option 2: *X* and *Y* cooperate, whilst *Z* defects: 2 points

Option 3: *X* and *Z* cooperate, whilst *Y* defects: 2 points

Option 4: All players cooperate: 3 points

Option 5: *X* defects, whilst *Y* and *Z* cooperate: 2 points

[19] This was William Ophuls' suggested solution for the tragedy of the commons: W. Ophuls, 'Leviathan or Oblivion', in *Toward a Steady State Economy* (San Francisco, CA: Freeman, 1973), pp. 215-30. For more on this solution to the tragedy of the commons, see, for example: Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action (Political Economy of Institutions and Decisions)* (Cambridge: Cambridge University Press, 1990), pp. 8-11.

Option 6: *X* and *Y* defect, whilst *Z* cooperates: 1 point

Option 7: *X* and *Z* defect, whilst *Y* cooperates: 1 point

Option 8: All players defect: 0 points

Observe now that the dominant strategy for *X* is to cooperate, with the same thing applying in *Y* and *Z*'s case. This is also the Pareto-optimal outcome. Thus, *X*, *Y*, and Z achieve the optimal equilibrium, avoiding the undesirable result of terrible traffic congestion (and pollution).

There are already familiar examples where a penalty is imposed by a Leviathan in order to collectively motivate or enforce a Pareto-optimal outcome, so this is not a novel idea. For example, the law already imposes a penalty (in the form of custodial sentences, fines, and so on) for certain kinds of non-cooperative behaviours towards others (e.g. theft, assault, murder, and so on). And egoists** should generally endorse this, on the basis that enforced mutual cooperation of this sort will very likely best serve their own individual self-interests**. Yes, they must forgo any possible benefits that might accrue to them from being able to act non-cooperatively towards others, but they then accrue the generally outweighing benefits of being protected from others doing the same to them. Likewise for the case at hand, where people would have to forgo the benefits of being able to drive to work whenever they want. Yet they would then accrue the outweighing benefits of not having terrible traffic congestion and pollution, thereby leading to an outcome that better serves each agent's own individual self-interest**.

In accord with H.L.A. Hart's idea of 'internal' and 'external' attitudes, whilst some form of measured punishment applied by the state may initially be required in order to prevent people from defecting (because they have an 'external' attitude that weighs the potential costs to the perpetrator of breaking the rules against the possible gains to them of doing so), over time most people can come to generally adopt a disposition in which they obey the rules without even thinking about the threat of punishment (because they have then adopted an 'internal' attitude that tends to unquestioningly obey the rules, and to see them as a constraint on their behaviour).[20] By such means, there is a shift of social norms, whereby people, as social animals, looking for signals from others about what is socially acceptable behaviour, tend to conform collectively to the newly prescribed behaviour.[21] And this only adds to the penalty for defection.

So, rationally, it seems that I ought to want such a penalty to be enforced, and for corresponding social norms to shift, on the basis that the outcome of this is better for me, in terms of better serving my own self-interest**. Of course, my self-interest** would be even better served by there being no penalty, everyone else cooperating, but me defecting. However, this is not a real-world option, because it is only by having the penalty that (almost) universal cooperation will be achieved.

As with any penalty-based scheme to motivate cooperative behaviour (e.g. laws and legal punishments), universal cooperation will never be achieved. However, the scheme would tolerate a degree of defection (with the defectors harming their own self-interests**) before it would become rational for everyone to start defecting. This is just as with society in general, where it is still generally in one's self-interest** to obey the law, despite the fact that some others (i.e. criminals) will not (with it only

[20] H.L.A. Hart, *The Concept of Law*, 3rd edn (Oxford: Oxford University Press, 2012 [1961]).
[21] E.g. Michael Hechter and Karl-Dieter Opp, *Social Norms* (New York: Russell Sage Foundation, 2005).

being in something like a state of anarchy that it would become rational for everyone to adopt a general policy of non-cooperation).

As such, because defection will also be the dominant strategy for everyone else, my choice is between having no penalty and (almost) everyone (including me) defecting, or having a penalty and (almost) everyone (including me) cooperating. (I have already shown why the option of me defecting if there is a penalty and others cooperate is not in my self-interest**.) The latter would better serve my self-interest** than the former, so it would appear to be rational for me to endorse the kind of scheme that would collectively motivate and enforce that mutual cooperation.

An alternative approach to the Leviathan that might work in certain circumstances (e.g. in a smaller and more stable community with an adequate social network) is for the people themselves to make a binding commitment to cooperative action, and to internally police and penalise defection (or hire a private agent as the enforcer). Elinor Ostrom proposes this as a possible solution to the tragedy of the commons, and presents known examples of it working in practice.[22] Thus, it may be the case that the optimal equilibrium (i.e. cooperate, cooperate, cooperate) can sometimes be achieved even without a Leviathan.

Whether enforced by an external Leviathan or within the group itself, it is in *X*, *Y*, and *Z*'s individual self-interest (on a self-interest** conception) to agree to abide by and promote some scheme whereby their car usage is restricted to collectively optimal levels and defectors are punished, since doing so enables them to achieve the Pareto-optimal equilibrium. As such, on egoism** this is then what they morally ought to do.

As with the cooperative solution to Baier's argument, it may again be objected that the numbers have been 'fixed' in order to get the right outcome, and that many

[22] Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action (Political Economy of Institutions and Decisions)*, pp. 9-18.

agents do not act rationally. Moreover, how do we know the ‘right’ outcome anyway? However, the point is that the magnitude of the penalty should be set such that the disutility of defection will always outweigh its utility in terms of serving agents’ self-interests**. So, to turn things around, I might say, yes, we *are* ‘fixing’ the numbers in order to produce the right result, because that is the very objective of the scheme.

In terms of agents not acting rationally, I think that there is a genuine worry to be had about agents defecting, even though doing so would not best serve their self-interest**. (Of course, contra my simplifying assumption earlier, the Leviathan will not always be able to detect defectors and impose the corresponding penalty, but that is to some significant extent a technological and logistical issue, and so is one that is in principle tractable.) However, based upon the case with law breaking in general, I would suppose that, even if there is always some residual defection, the presence of the penalty (which can be adjusted up or down over time, until the optimal level is found), together with the corresponding social norms, will ensure that most people will cooperate most of the time.

In terms of the last objection, whether the rigid enforcing of mutual cooperation, the unrestricted permitting of defection, or something in between, is the ‘right’ outcome may not be entirely clear. Some cases, such as the aforementioned penalties against murder are clear-cut, but others are rather less so. *Ex hypothesi*, it seems that the first outcome will be the best one, and so in one sense that is our answer (and is why I am taking that to be the ‘right’ outcome here). However, this may not accurately reflect the real world. What is required is to determine which outcome, on balance, will best serve the self-interest** of most individual agents. This is a complex question on many levels, and so something that I shall set aside.

Thus, returning to Argument 11, I would argue that it therefore fails. Specifically, conclusion C1 would no longer follow from premises P1, P2, and P3, because we may (and, from a self-interest** perspective, should) attach a penalty to driving, even if *X* prefers to drive, and driving into work will have only an infinitesimal effect upon the overall volume of traffic (on the basis that not driving better serves *X*, *Y*, and *Z*'s self-interest**). I view this not as a purely paternalistic move (like levying a penalty upon people smoking in the privacy of their own homes, on the basis that doing so is in their individual self-interest**), because, in this case, defecting has a direct impact upon *other* people.

Consequently, I would argue that Goal Theory's particular variant of egoism survives the challenge that it is collectively self-defeating. Moreover, based upon my earlier analysis, it is also appears to provide solutions to conflicts of interest, and avoid charges of being self-contradictory. On this positive note, I shall turn now to external criticisms of egoism.

5.4 External criticisms

As explained earlier in section 5.1, the general form for an external criticism of Goal Theory's particular variant of ethical egoism may be represented by the following modus tollens argument:

Argument 12

P1)	Egoism** implies proposition *p*.
P2)	Proposition *p* is false.
C)	Therefore, egoism** (and thence Goal Theory) is false.

In this section, I shall evaluate a paradigmatic case (and reference several others) that can be framed in terms of some putative *p* that is supposedly implied by egoism**, but where *p* is allegedly false (thereby entailing the apparent falseness of egoism**, and thence Goal Theory). In response, I shall propose the following three possible moves (equivalents of which are commonly made by utilitarians): (1) rejecting premise P1, by denying that egoism**, when properly understood and applied, actually implies the *p* in question; (2) rejecting premise P2, by 'biting the bullet' — accepting that egoism** may imply *p* in certain circumstances, but then denying that *p* would then be false *in those circumstances*, no matter how counter-intuitive that may be; and (3) granting premises P1 and P2 in the circumstances specified, but arguing that this would not arise in the *real* world, only in some other possible one, yet Goal Theory (and thence egoism**) is only designed to provide moral guidance in the real world.

As a paradigmatic example of a criticism of ethical egoism that takes the above form, consider a hypothetical case presented by Fred Feldman (and cited by Keith Burgess-Jackson). Feldman maintains that his example effectively refutes egoism, setting it out as follows:

> A man is the treasurer of a large pension fund. He is entrusted with the job of keeping track of and investing the money deposited by the workers. When a worker retires, the

> worker is entitled to draw a weekly sum from the fund. Suppose the treasurer discovers that it will be possible for him to use all the money for his own selfish pleasure without being caught. Perhaps he wants to buy a large yacht and sail to a South Sea island, there to live out his days in idleness, indulgence, procreation, and, in a word, enjoyment. Since there is no extradition treaty between the South Sea island and the United States, he can get away with it. Let us also suppose that if the treasurer does abscond with the funds, hundreds of old people will be deprived of their pensions. They will be heartbroken to discover that instead of living comfortably on the money they had put into the pension fund, they will have to suffer the pain and indignity of poverty.[23]

Feldman thinks that, on egoism, stealing the money would be the right thing to do. Yet he claims that this is *not* in fact the right thing to do. As such, and though he thinks egoism can be formulated consistently, Feldman proclaims it to be false.[24] Expressed syllogistically (and substituting egoism** in particular for egoism in general):

Argument 13

P1)	Egoism** implies that the treasurer of the pension fund ought to steal the money.
P2)	It is false that the treasurer of the pension fund ought to steal the money.
C)	Therefore, egoism** (and thence Goal Theory) is false.

In line with the first move proposed earlier, I would challenge premise P1, arguing that egoism**, when properly understood and applied, very probably does *not* imply that the treasurer ought to steal the money. Accordingly, in failing to correctly identify or to adhere to what will best serve his strongest desire over the long-term, the treasurer

[23] Fred Feldman, *Introductory Ethics* (London: Pearson, 1978), p. 95.
[24] Concurring with this evaluation we find: J.W. Cornman, K. Lehrer, and G.S. Pappas, *Philosophical Problems and Arguments: An Introduction*, 3rd edn (New York: Macmillan Publishing Company, 1982).

would be making a moral mistake on my account by stealing the money. Perhaps stealing the money might best serve the treasurer's strongest *present desire* at some particular time. After all, as Feldman implies, it is conceivable that the treasurer's strongest present desire at some particular time is for a large yacht, and stealing the money will allow him to buy that yacht. In that case, stealing the money would be commanded on egoism*.

However, I find it highly improbable that stealing the money would best serve the treasurer's strongest *enlightened* desire over the *long* term (and so it very probably would *not* be endorsed on egoism**, and thence Goal Theory). On the assumption that the thing that the treasurer would desire most, if he was fully rational and sufficiently informed, is something in the region of the universal true strongest desire that I adduced in section 2.5, viz. a kind of deep and abiding satisfaction, then I would argue that there are good reasons to think that stealing the money would not best serve this desire over the long term.

First, if those in the treasurer's new community became aware of his theft, then he would probably acquire a bad reputation, meaning that others would not trust him and would be disinclined to cooperate with him (viewing him, in game-theoretic terms, as a 'cheater', and thus not someone that they would want to employ, go into business with, or even be neighbours or friends with). As a result, he would probably find himself being shunned, and he would likely forego (at least some of) the rewards of direct and indirect reciprocity. He might even face retribution from those whose pension money he has stolen, or from others motivated to act on their behalf. And to the extent that he engaged in lies and cover-ups in order to avoid others learning about what he has done, then the strain of maintaining a consistent web of lies, as well as the constant fear of being unmasked, would generate anxiety. These kinds of bad consequences stand in opposition to him best serving over the long term anything like the kind of strongest enlightened desire proposed.

Second, he runs the very real risk of eventually being apprehended and brought to justice. Even if there is currently no extradition treaty, one might be established in the future. Moreover, there might be other outweighing reasons for him to return home (e.g. access to health care, to be with family, and so on). This is what eventually happened with Ronnie Biggs. Upon his return to the UK, he was immediately arrested and imprisoned. In accord with my argument, he is reported to have said that 'It has not been an easy ride over the years. Even in Brazil I was a prisoner of my own making. There is no honour to being known as a Great Train Robber. My life has been wasted.'[25] If this were to happen in the treasurer's case, then he would likely face imprisonment, which would almost certainly not be conducive to him best serving over the long term anything like the suggested strongest enlightened desire.

Finally, by being forced to live with the knowledge that 'hundreds of old people will be deprived of their pensions' and 'will be heartbroken to discover that instead of living comfortably on the money they had put into the pension fund, they will have to suffer the pain and indignity of poverty', the treasurer is likely to feel some degree of guilt, shame, remorse, disappointment, self-loathing, and so forth. These kinds of emotional costs are antithetical to him best serving over the long term anything like the sort of strongest enlightened desire that I put forward.

Now, it might be said that, to the extent that these emotions do not reinforce behaviour that is in the treasurer's own self-interest**, then they are irrational, and he should try to overcome them. I would agree with that claim, and from my earlier analysis of the evolutionary origin of our moral emotions, it is clear that I think moral

[25] Jonathan Owen and Sadie Gray, 'Ronnie Biggs Pleads: Let Me out So I Can Die with My Family', *Independent on Sunday*, (30th December 2007).

emotions will not always align with what Goal Theory demands. In those cases, we should certainly try to overcome them, so that they do not divert us from best serving our true strongest desire over the long term. However, there are many cases where they do align, and emotions such as guilt, shame, and remorse probably evolved in order to reinforce the very kinds of cooperative behaviour that Goal Theory does generally demand, and which the treasurer would have failed to exhibit. Thus, these sorts of emotions will often not be irrational from Goal Theory's perspective. They might not be an *independent* reason to act in a certain way (since they *can* be irrational), but where, as in this case, they align with the first two reasons given, then they function as a legitimate additional impediment to an uncooperative agent best serving their true strongest desire.

So, if doing *x* in *C* will already tend to frustrate agent *A*'s true strongest desire (for the previous two reasons), then an emotion (such as guilt or shame) that aligns with Goal Theory would be one that opposes *A* doing *x* in *C*, thereby reinforcing the other reasons for *A* to not do *x* in *C*. As such, these emotions would be rational from a Goal Theory perspective, and ones that ought, I suggest, to be cultivated, so that they come easily and we are more likely in general to heed them (whilst remaining mindful that there may be exceptions to be vigilant for).

By contrast, if he does not steal the money, then the treasurer need suffer no consequent guilt, shame, disappointment, self-loathing, and so on. Moreover, he may cultivate a good reputation, gain the trust of others, more reliably benefit from their direct and indirect reciprocity, avoid being ostracised, and escape their retribution. This need not require a great deal of time or effort — it would be sufficient to just routinely obey the law, act with honesty and integrity, show some compassion, and suchlike. Lastly, he no longer runs the risk of being apprehended by the law for the theft, and facing the consequent punishment. All of these are conducive to him best serving the kind of strongest enlightened desire suggested. Of course, in not stealing the money, the treasurer does then forgo the potential excitement that might come from his new life, as well as the easy access to luxuries and suchlike. However, it is my contention that these benefits would be of a kind that (at most) would only tend to best serve the treasurer's strongest present desire at a particular time, rather than best serving his strongest enlightened desire over the long term (and, as such, aligning more with egoism* than egoism**).[26] Thus, with the consequences of stealing the money appearing to so much less reliably serve the treasurer's strongest enlightened desire over the long term than not stealing the money would do, I would argue that egoism** very probably does *not* imply that the treasurer of the pension fund ought to steal the money.

Incidentally, this sort of reasoning explains why egoism** is probably immune to some standard objections to utilitarianism, including its difficulty in accounting for supererogatory actions and the obligatoriness of promises. In the former case, there are both psychological payoffs and benefits in terms of direct and indirect reciprocity to being such a 'supercooperator', with these plausibly translating into the enhanced serving of one's strongest enlightened desire over the long term. Likewise in the latter case, where cultivating a reputation as someone who does not break (even inconvenient) promises will likely garner greater cooperation from others, to the benefit of one's own self-interest**.

In principle, there might be exceptions to what I have argued, and perhaps the treasurer constitutes just such an exception. So, what ought he to do? Well,

[26] Ronnie Biggs' remarks support this contention. Moreover, there is good evidence that additional money beyond moderate levels (≈ $75,000 in the US in 2010) brings little benefit in terms of extra happiness: E.g. Daniel Kahneman and Angus Deaton, 'High Income Improves Evaluation of Life but Not Emotional Well-Being', *Proceedings of the National Academy of Sciences*, 7 (2010), 16489–93.

remembering Epicurus, one can never reliably know that one really will be an exception, able to treat others badly yet still escape a miserable outcome. Therefore, when an agent knows that a certain kind of conduct is in generally in their best self-interest** (e.g. not stealing), and they do not know that they constitute a genuine exception to this, then I submit that they ought to act accordingly. (The 'rule' here may not be as simple as 'do not steal from others', but might instead be of the form: 'do not steal from others, *except* when you know that stealing will best serve your true strongest desire.') From thoughts like this, we may derive a kind of 'rule-egoism', stated by Gregory Kavka as follows:

> Each agent should attempt always to follow that set of general rules of conduct whose acceptance (and sincere attempt to follow) by him on all occasions would produce the best (expected) outcomes for him.[27]

I would tentatively endorse a 'rule' modification to egoism** to apply in any cases where we can establish as a general rule that agents ought do *x* (e.g. not steal money), on the basis that doing *x* will generally best serve their strongest enlightened desire over the long term, and they do not know that they, in the circumstances in which they find themselves, constitute a genuine exception to this rule. As Burgess-Jackson observes, few have considered the possibility that egoism may be modified in this way (as utilitarianism has been).[28]

This does not change the basic theory. One still ought to do what will best serve one's strongest enlightened desire over the long term. However, in practice, this may sometimes not be ascertainable in anything like real-time. As such, for *pragmatic*

[27] G.S. Kavka, *Hobbesian Moral and Political Theory* (Princeton, NJ: Princeton University Press, 1986), pp. 358-59.

[28] Though Kagan has: Shelley Kagan, *Normative Ethics* (Boulder, CO: Westview Press, 1998), pp. 194-204.

reasons, if one does not know that one constitutes an exception to a general rule (e.g. that one is an exception to the general rule to tell the truth, because one needs to deceive an enquiring murderer), then one should probably follow the general rule (knowing that, in most cases, doing so will best serve one's strongest enlightened desire over the long term). This relates to what I said in section 2.3 about the value of *approximate* knowledge. Even if we do not know the right thing to do, we can still know what the right thing is, given what we know so far. And that is optimal in the absence of perfect knowledge. Thus, we might not be able to determine in a suitable timeframe what will best serve our strongest enlightened desire in some specific circumstances, but we can still know what will best serve the strongest enlightened desires of *most* agents in *most* circumstances, and then adopt this as a general rule.

Following on from this, I would submit that it will likely be beneficial for the egoist** to cultivate appropriate habits of character that will make the above-mentioned kind of rules of conduct easy to perform, happening without the necessity of conscious effort. If we call these habits of character 'moral virtues' (e.g. integrity, compassion etc.), then I am suggesting that the egoist** should generally cultivate those virtues, in order that the corresponding sorts of behaviour are more reliably produced and alternative kinds of behaviour more reliably avoided. Once these virtues are habituated, then moral agents will more reliably and readily act accordingly, which will generally be in their own self-interest**. Thus, from egoism**, we may derive a kind of virtue ethics (on which the egoist has *proximate* reason to cultivate the virtues in question, with the *ultimate* reason being to best serve his or her self-interest**).

Returning to the thought experiment, Feldman might be prepared to grant that, on egoism**, most people most of the time ought not to steal the pension money. However, he may still object that there are plausible exceptions to this. Some agents can effectively act with impunity (towards everyone, or merely towards those outside their circle of family and acquaintances[29]). And, for these agents at least, Feldman might maintain, egoism** *does* imply the kind of *p* in question. After all, if an agent possesses absolute power within some domain, and so is effectively untouchable, then why should they not act exactly as they want, paying no heed to the interests of anyone else. (This is an ancient problem, going back at least as far as the Gyges myth in Plato.) Would egoism** not dictate that they do precisely that?

There are responses that I might make here. Despite their easy access to immense power and abundant material goods, despots often seem to live paranoid lives full of anger, disappointment, and fear, leavened only with relatively fleeting and hollow forms of happiness, rather than deeper and more abiding kinds of happiness, satisfaction, and contentment. This is because they do not do (or do far less of) the sort of things that reliably bring people deeper and more abiding kinds of happiness and satisfaction (e.g. acts of compassion and compersion, expressing gratitude and kindness, and cultivating strong social bonds), and they do far more of the things that reliably bring people misery (e.g. acts of cruelty and violence that increase the risk of reprisals and assassination attempts, as well as the fear of these).[30] Although having power is generally correlated with well-being (and vice versa), since one is able to live a more self-determined life, striving for power is not (and one can find many other,

[29] Here I shall analyse the former, taking the latter to be a half-case that suffers the same problems.

[30] On this, see, for example: J. A. Piliavin, 'Doing Well by Doing Good: Benefits for the Benefactor', in *Flourishing: Positive Psychology and the Life Well-Lived*, ed. by C. L. M. Keyes and J. Haidt (Washington, DC: American Psychological Association, 2003), pp. 227– 47; P. A. Thoits and L. N. Hewitt, 'Volunteer Work and Well-Being', *Journal of Health and Social Behavior,* 42 (2001), 115– 31; N. Weinstein and R. M. Ryan, 'When Helping Helps: Autonomous Motivation for Prosocial Behavior and Its Influence on Well-Being for the Helper and Recipient', *Journal of Personality and Social Psychology,* 98 (2010), 222– 44. Also: R. F. Baumeister and M. R. Leary, 'The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation', *Psychological Bulletin,* 117 (1995), 497–529; Lorraine Besser-Jones, 'Personal Integrity, Morality and Psychological Well-Being: Justifying the Demands of Morality', *Journal of Moral Philosophy,* (2008), 361–83.

less self-destructive, ways to live a more self-determined life than by becoming a despot).[31]

What is more, a fully rational and sufficiently informed despot would have to agree that their subjects ought to kill them — raising the question of why a person would choose to be someone they admit deserves to be killed (and if they did, by what means they could find deep and lasting happiness and contentment, knowing they are the sort of person who ought to be killed). Besides, the history of actual despots does not support statistically good odds of that going well for them over the long term, with a significant minority being murdered, executed, committing suicide, dying in exile, or being brought to justice. For example, according to noted 'atrocitologist' Matthew White, of those dictators who have died, around 40% were murdered or executed, committed suicide, or died in war, prison, or exile.[32] For all of these reasons, I would argue that, even in the case of a person who can apparently act with impunity, egoism** very probably would not imply that they lie, steal, murder, and so on.

However, if, for the sake of argument, I was to grant that egoism** may in certain exceptional circumstances imply that the treasurer of the pension fund ought to steal the money, then how would I respond? Well, in such a hypothetical case, I would make the second move described earlier — biting the bullet by accepting this, but then arguing that, *in those circumstances*, stealing the money really *would* be what the treasurer morally ought to do. In other words, I would then reject the minor premise,

[31] See: Yona Kifer and others, 'The Good Life of the Powerful: The Experience of Power and Authenticity Enhances Subjective Well-Being', *Psychological Science*, 24 (2013), 280-88.

[32] See: Matthew White, *The Great Big Book of Horrible Things: The Definitive Chronicle of History's 100 Worst Atrocities* (New York: W. W. Norton & Company, 2011). Of those murdered or executed, we have, for example, Benito Mussolini, Nicolae Ceausescu, Saddam Hussein, and Moammar Gaddhafi. Those committing suicide include Adolf Hitler. Those brought to justice include Slobodan Milosevic and Hosni Mubarak (though he was eventually acquitted). And those dying in exile include Jean-Claude Duvalier, Mobutu Sese Seko, Ferdinand Marcos, and Idi Amin. Pol Pot is also suspected of either committing suicide or having been murdered.

i.e. P2, of Argument 13. This kind of bullet biting is common amongst utilitarians. For example, as JJC Smart says:

> Admittedly utilitarianism does have consequences which are incompatible with the common moral consciousness, but I tended to take the view ‘so much the worse for the common moral consciousness’. That is, I was inclined to reject the common methodology of testing general ethical principles by seeing how they square with our feelings in particular instances.[33]

As I said in section 3.4, Peter Singer also adopts this approach, by undertaking a top-down approach to moral theorising, whereby one selects ‘a theory that is based on a fundamental axiom that seems ... clear and undeniable’, and then applying the theory to particular situations, biting whichever bullets are then implied. This bullet-biting approach is routinely deemed legitimate when employed by utilitarians such as Singer and Smart. Yet, as observed by Burgess-Jackson, by an apparent double standard it is sometimes prohibited when egoists do the same.[34]

How might the proponent of the argument expressed in Argument 13 attempt to press the case for premise P2 (i.e. it is false that the treasurer of the pension fund ought to steal the money)? I suggest a standard move would be to invoke moral intuition. Modifying Argument 13, by making fully explicit this defence of premise P2, we would have:

[33] JJC Smart, 'An Outline of a System of Utilitarian Ethics', in *Utilitarianism: For and Against*, ed. by JJC Smart and B. Williams (Cambridge: Cambridge University Press, 1973), pp. 1-74 (p. 68).

[34] Burgess-Jackson, 'Taking Egoism Seriously', p. 539. He cites some flagrant examples of this apparent double standard from James Rachels and William Shaw.

Argument 14

P1)	Egoism** implies that the treasurer of the pension fund ought to steal the money.
P2)	According to our moral intuition, it is false that the treasurer of the pension fund ought to steal the money.
P3)	One is justified in believing on the (sole) basis of a putative source of evidence only if one lacks (undefeated) reason to think it unreliable.
P4)	We lack (undefeated) reasons to think moral intuitions are unreliable.
C1)	Therefore, we are justified in believing on the (sole) basis of moral intuitions.
C2)	Therefore, it is false that the treasurer of the pension fund ought to steal the money.
C3)	Therefore, egoism** (and thence Goal Theory) is false.

However, I think that the sub-argument from P2 to C2 fails as a defence of premise P2 of Argument 13. Specifically, I would now reject the new premise P4. As discussed in some depth in section 3.4, I think we *do* have (undefeated) reasons to think moral intuitions are unreliable, and we have no generally accepted means to distinguish any trustworthy intuitions from untrustworthy ones. However, if P4 is probably false, then conclusion C2 is not shown to be true; and that leaves premise P2 of Argument 13 unproven.

Moreover, based upon my positive argument in section 2.1 (and my subsequent analysis), in Goal Theory (and thence egoism**) I think we have a theory that is coherent and at least *prima facie* plausible, albeit defeasible. Thus, if ever premise P1 of Argument 13 were true in some exceptional real-world circumstances, and this result were to clash with our moral intuition, then I would have more faith in egoism** than in any conflicting moral intuition. This would not be *blind* faith, however, but what I would argue is a *justified* faith, based upon an evaluation of the relative epistemic merits of Goal Theory (and thence egoism**) and moral intuition. Hence, I submit that conclusion C2 of Argument 14 would not merely be unproven, but probably false; and that would leave premise P2 of Argument 13 probably false too.

Now, Feldman may object that it is *his* thought experiment, and so he gets to stipulate that the circumstances are such that egoism** (when properly understood and applied) *does* imply that the treasurer of the pension fund ought to steal the money (per premise P1 of Argument 13), even if this would not be so for almost all people in almost all circumstances; and that it is *false* that the treasurer of the pension fund ought to steal the money (per premise P2). However, mere stipulation does not make something true in the real world. Accordingly, by recourse to the third of the moves presented earlier, I would argue that, in making such stipulations, Feldman effectively renders his thought experiment a fanciful hypothetical, thereby relocating it from the real world to some other possible world instead. Yet Goal Theory (and thence egoism**) is designed to provide moral guidance in the *real* world, and not necessarily in other possible worlds, so this need not be of any real concern to me.

In conclusion, I reject Argument 13. First, I argued that egoism**, when properly understood and applied, very probably does not imply that the treasurer ought to steal the money — contra premise P1. Second, if ever it did imply this in some non-fanciful, real-world scenario, then I would deny premise P2 *in those circumstances*, on the basis that we should have more justified faith in egoism** than in any conflicting moral intuition. If Feldman wants to refute this conclusion, then he must: (1) identify some real-world circumstances in which egoism**, when properly understood and applied, really would imply that that the treasurer of the pension fund ought to steal the money; and (2) demonstrate that we should have more justified faith in the particular moral intuition with which this result clashes than in Goal Theory (and thence egoism**), notwithstanding what I have argued to the contrary. In the meantime, I submit that my account is not undermined by Argument 13.

Feldman suggests that we may construct many other, similarly decisive, examples to that of the pension fund treasurer. I agree, but since I find the treasurer example to be not at all decisive, then, by implication, I find these other examples to be similarly unpersuasive.

In general, I would argue that egoism**, when properly understood and applied, will rarely if ever generate real-world propositions that conflict with our stock moral truisms (where conflicting with these truisms would be commonly regarded as implying that the propositions in question are thereby false, on the basis that stock moral truisms are, by definition, widely accepted and intuitively true). This means, for example, that, in the real world, I think egoism** would rarely if ever command a moral agent to commit murder, to steal from a pension fund, or to order genocide (to reference three examples previously discussed). However, if ever it did, then, for the reasons explained, I would have more justified faith in the correctness of egoism** than in the conflicting stock moral truism. (Egoism**, and thence Goal Theory, is still *defeasible* though — just not on the basis of unsupported moral intuition.)

In general terms, the criticism of egoism is that it implies we have no moral duties to other people. Substituting egoism** in particular for egoism in general, we might express this criticism as:

Argument 15

P1)	Egoism** implies that agents do not have certain moral duties to other people.
P2)	But agents do have these moral duties to other people.
C)	Therefore, egoism** (and thence Goal Theory) is false.

On egoism**, the only underived moral duty we have is to ourselves, insofar as we each ought to do what will best serve our own strongest enlightened desire over the long term. However, the egoist** would challenge premise P1, on the basis that, for contingent (e.g. emotional, social, and legal) reasons, doing what will best serve our own strongest enlightened desire over the long term entails a general obligation upon us to act in certain paradigmatically ‘moral’ (e.g. honest, compassionate, and altruistic) ways towards others. Accordingly, we acquire certain contingent, derived moral duties towards others. Thus, for any plausible moral duty towards others, *d*, that the critic might adduce (e.g. a duty to rescue someone from drowning, or a duty to not steal a pension fund’s money), the egoist** would claim that, when properly understood and applied, egoism** very likely generates *d* too, but as a *derived* moral duty. Hence, they would argue that, in (almost) all plausible, real-world circumstances of the sort considered, premise P1 would be false.

However, if egoism** ever failed to generate some particular *d*, then, in those circumstances, the egoist** would have more faith in the correctness of their theory than in any conflicting intuition, arguing that there is therefore no such moral duty in those circumstances. By such means, egoism** would align with and vindicate certain moral duties (accounting for and justifying the basis of these duties, rather than leaving them ultimately mysterious or dependent upon intuition), whilst perhaps eliminating others (and providing good grounds for this elimination).

One putative moral duty that egoism is often supposed to not generate is that of self-sacrifice (e.g. a parent dying in order to protect a child, or a soldier falling on a grenade for their comrades), since there is no long-term gain to outweigh the short-term loss. However, egoism** may generate at least some of these duties, on the basis that our true strongest desire might be better served over the long-term by dying rather than having a lifetime of regret and self-loathing or other misery (i.e. no life may be better on balance that a life of misery). And if it did not generate a particular *d*, then I would deny that there is such a duty.

5.5 Egoism is unacceptably arbitrary

Let me now turn to a different argument that has been directed at egoism. According to James Rachels, this is the argument that comes closest to an outright refutation of egoism. He writes that egoism:

> …advocates that each of us divide the world into two categories of people – ourselves and all the rest – and that we regard the interests of those in the first group as more important than the interests of those in the second group. But each of us can ask, what is the difference between myself and others that justifies placing myself in this special

> category? Am I more intelligent? Do I enjoy my life more? Are my accomplishments greater? Do I have needs or abilities that are so different from the needs or abilities of others? What is it that makes me so special? Failing an answer, it turns out that Ethical Egoism is an arbitrary doctrine, in the same way that racism is arbitrary.[35]

Expressed more formally:

Argument 16

P1)	Any moral doctrine that assigns greater importance to the interests of one group than to those of another is unacceptably arbitrary unless there is some difference between the members of the groups that justifies treating them differently.
P2)	Ethical Egoism would have each person assign greater importance to his or her own interests than to the interests of others. But there is no general difference between oneself and others, to which each person can appeal, that justifies this difference in treatment.
C)	Therefore, Ethical Egoism is unacceptably arbitrary.

Here I would reject premise P2. Translating into the language of egoism**, in order to direct the argument at my account specifically, the charge would be that egoism** would have each person assign greater importance to his or her enlightened self-interest than to the enlightened self-interests of others, when there is no general difference between oneself and others, to which each person can appeal, that justifies this difference in treatment.

[35] Rachels, 'Ethical Egoism', p. 199.

In order to capture adequately the broad kind of self-interest being appealed to here, let me define the following:

> **Self-interest*****: doing *F* is in an agent's self-interest*** if it will result in some of their enlightened desires being satisfied.[36]

Now, I would submit that there *is* a general difference between me and other people, to which I can appeal, that justifies me assigning greater importance to my own self-interest*** than to the self-interest*** of others, viz. I am *me*, and other people are not.

To see why this matters, imagine that, in some circumstances *C*, agent *A* can do *x* or ~*x*. Imagine further that doing *x* in *C* will serve *A*'s own self-interest***, but doing ~*x* will only serve the self-interest*** of others. Now, from the definition of self-interest***, if doing *x* in *C* will serve *A*'s self-interest***, then doing *x* in *C* will result in some of *A*'s enlightened desires being satisfied. By reference to the HTR* (section 2.6), this means that *A* will then have *pro tanto* normative reason to do *x* in *C*. By contrast, if doing ~*x* serves only the self-interest*** of others, not serving *A*'s self-interest*** at all, then this means that doing ~*x* in *C* will then result in *none* of *A*'s enlightened desires being satisfied. Thus, from the HTR*, *A* will *not* then have *pro tanto* normative reason to do ~*x* in *C*. Thus, I think *A* may legitimately assign greater importance to doing *x* in *C* than to doing ~*x* in *C*, on the basis that they have a *pro tanto* normative reason to do *x* in *C*, but no *pro tanto* normative reason to do ~*x* in *C*.

[36] My earlier conception of self-interest**, on which doing *F* is in an agent's self-interest** if it will result in their strongest enlightened desire being satisfied over the long term, is appropriate within the context of a definition of egoism**, but is too narrow for this purpose, since an action may not serve an agent's self-interest**, yet still satisfy some of their enlightened desires.

In other words, on my account, satisfying my own enlightened self-interest will satisfy some of my enlightened desires, meaning that I will then have normative reason to do this (on the HTR*). By contrast, satisfying only the enlightened self-interest of others will *not* satisfy any of my enlightened desires, meaning that I will *not* have normative reason to do this (on the HTR*). Thus, I would argue that there is reason for me to assign greater importance my own enlightened self-interest than to the enlightened self-interest of others, with this reason deriving from whose enlightened desires are being satisfied in each case. (Of course, I am not suggesting that this applies only to *me* — the same applies with other agents and their own enlightened self-interests.)

5.6 Conclusions

In this chapter, I wanted to answer the charge that ethical egoist theories — on which agents ought to do what best serves their own self-interest — are defeated by a conjunction of internal and external criticisms (meaning that Goal Theory, with its egoistic account, is fatally undermined).

At the beginning of this chapter, I pointed out that egoism accrues a number of advantages over rival theories, including that it avoids any possible conflict between self-interest and morality, that agents have a ready answer to why they should be moral, and that it makes moral behaviour rational by definition (on the assumption that it is rational to pursue one's own interests). The critic of egoism might be prepared to grant some or all of these advantages, but would then argue that it suffers from outweighing disadvantages, perhaps including that it is collectively self-defeating, cannot deal with conflicts of interest, is logically inconsistent, is unacceptably arbitrary, and implies moral propositions that are false.

In this chapter, I have examined these objections to egoism, finding that whilst some may defeat simplistic and primitive versions of egoism (such as the view that I call egoism*, on which agents ought always to do what will best serve their strongest unenlightened desire over the short-term), none appears to trouble Goal Theory's more sophisticated form of egoism (which I label egoism**). Specifically, it seems that, on game-theoretic grounds, egoism** *can* provide solutions to conflicts of interest, and is *not* collectively self-defeating. Nor does it appear to be self-contradictory or unacceptably arbitrary. Moreover, I showed that when egoism** is properly understood and applied, it probably does *not* imply the *prima facie* false propositions adduced; and if ever it did in some not entirely fanciful real-world circumstances, then it would be too bad for any intuition to the contrary. Thus, egoism** appears to accrue the above-mentioned benefits of egoism, whilst simultaneously resisting the standard objections.

Once again, I would submit that my account captures enough of the appearances insisted upon by non-egoists to be, overall, plausible. In particular, because I suggest that cooperation will generally be conducive to serving agents' strongest enlightened desires over the long term (and vice versa), then I maintain that my account will generate (almost) all of the standard moral duties to others (to help them, to treat them honestly, to be compassionate towards them, and so on), with few, if any, genuine, real-world exceptions (notwithstanding the *prima facie* implications of thought experiments such as Feldman's). These duties will not be *underived* ones, however. With the exception of the underived moral duty we have to *ourselves* (insofar as we each ought to do what will best serve our own strongest enlightened desire over the long term), I deny that there are such things. Rather, they will be *derived* duties. Moreover, although on my account there is no requirement for agents to give weight to the interests of others *per se*, I submit that agents will generally best serve their strongest enlightened desires over the long term by acting *as if* the interests of others do have independent weight (e.g. by not breaking inconvenient promises, on the basis that any short-term losses incurred will be outweighed by long-term gains, such as having one's future promises trusted).

As noted at the beginning of the chapter, egoism is not a popular theory, with many philosophers disliking it intensely. However, even then, some find it attractive (including Burgess-Jackson, Hobbes, Tibor Machan, John Hospers, Jesse Kalin, and Edward Regis[37]). And Sidgwick accords it the same status as utilitarianism when he says that 'the aim of furthering one's own interest stands on just as rational a basis as the aim of furthering the universal interest'.[38] In any case, the adjudication of whether a moral system is true is not a matter of mere consensus or majority belief. With egoism**, I submit that we have an account that survives the standard objections to egoism, and improves both upon less sophisticated egoist accounts and upon non-egoist accounts (which struggle to avoid conflicts between self-interest and morality, to supply us with reasons to comply with morality, and to give us the motivation to actually do so).

Now that I have answered all three of the dominant challenges that I identified in section 1.1, it is now time to return to considerations of theoretical adequacy. Accordingly, in the next chapter, I shall seek to establish if Goal Theory successfully meets all of the adequacy criteria against which I am assessing it.

[37] Burgess-Jackson, 'Taking Egoism Seriously', p. 540.
[38] Quoted in: Singer, 'Sidgwick and Reflective Equilibrium', p. 504.

Chapter 6

Assessing Goal Theory's theoretical adequacy

Now it is time for a reckoning. As explained in chapter 1, all of the familiar theories struggle to satisfy at least one of the theoretical adequacy criteria against which I am testing (in addition to facing the kind of serious objections described).[1] Why should Goal Theory (as a particular reductive naturalist account) fare any better in this regard? In this chapter, I aim to establish that it does, with it plausibly satisfying *all* of these criteria, including those with which naturalist accounts generally have difficulty.

To begin with, I shall review where Goal Theory stands thus far in relation to the conditions that are generally acknowledged to bear upon the theoretical adequacy of any metaethical theory that seeks (as Goal Theory does) to answer the basic ontological question: 'what is the nature of moral reality?' As a reminder, here are the criteria in question, as originally enumerated in section 1.1 (changing the order for convenience):

1. It would plausibly account for the supervenience of the moral world upon the non-moral one, such that it is impossible for the former to differ unless there is also a difference in the latter.

[1] My intention there was not to present an exhaustive taxonomy of metaethical positions. However, I did list the main metaethical positions in the landscape, based upon whether or not they posit objective moral facts (realism vs. antirealism), whether any such facts reduce to or otherwise fit with natural facts (naturalism vs. non-naturalism, and reductive naturalism vs. non-reductive naturalism), and, if there are no such facts, then whether or not moral statements nonetheless purport to state moral facts (error theory vs. expressivism).

2. It would have an adequate moral epistemology, accounting for how we can apprehend anything to be known within morality.
3. It would be ontologically parsimonious, in not multiplying entities beyond necessity.
4. It would be conservative, in preserving many of our existing moral beliefs that are widely held, supportive of other beliefs, and resistant to alteration after reflection.
5. It would explain why it is that, necessarily, anyone who sincerely holds a moral view is motivated to some extent to comply with it.
6. It would explain how moral requirements entail excellent reasons for compliance.
7. It would be able to account for the relatively greater depth and breadth of moral disagreement, as compared with other areas of supposed objective truth (where a failure to do this is argued to undercut a theory's claim to provide *objective* moral judgements).
8. Finally, it would have a semantics of moral discourse, supplying plausible answers to well-known semantic puzzles (e.g. Moore's Open Question Argument, and Horgan and Timmons' Moral Twin Earth experiment).

I would submit that I have effectively demonstrated Goal Theory's compliance with conditions 1 to 6 already. Though I shall not rehearse my arguments here, I think a recap would be useful.

Supervenience: as a reductive naturalist account, conceiving of moral facts and properties as being reductively identical to certain natural ones (composed of a particular conjunction of natural facts of cause and effect and idealised human desire), Goal Theory offers a straightforward account of the supervenience of the moral world on the non-moral one, avoiding the kind of problems that plague non-reductive accounts. Specifically, let *N* be a complete description of all of the natural facts and properties of an act, event, or situation. Then, if two acts, events or situations are *N*, we know that any natural facts and properties of true strongest desires and what best serves those desires will be the same. However, in that case, the two acts, events, or situations will also be identical on Goal Theory in all *moral* respects.

Epistemology: I explained in section 2.3 how, in locating the domain of morality within the familiar natural world, Goal Theory's moral facts and properties are naturalistic ones that are in principle discoverable by the familiar methods of science (rather than by appeal to some special faculty or other means by which we are supposed to apprehend non-natural *sui generis* facts and properties). The methods of discovery and justification may be complex and difficult, but there is much precedent here, as science has discovered and justified many things in the teeth of methodological difficulties (including in the fields of cosmology, particle physics, psychology, sociology, and cognitive science, for example), so I think there is good reason to imagine that the discovery of moral facts and properties will yield to a suitable research programme.

Even if we are never able to access perfect knowledge in this area, then *approximate* knowledge of the necessary human psychology and cause and effect (and thence of the relevant moral facts) is still valuable. Moreover, we may also be able to establish general rules to follow in cases where the complete calculations of which actions best serve particular agents' true strongest desires in such and such circumstances are too difficult, burdensome, or time-consuming to make. Accordingly, with moral facts on Goal Theory being in principle as accessible and epistemically secure as other natural facts, I argued that Goal Theory has an adequate moral epistemology.

Ontological parsimony: I argued in section 2.2 that, in reductively identifying moral facts and properties with established natural ones, Goal Theory is as ontologically parsimonious as moral error theories (positing the same natural facts and properties, and differing only from anti-realist theories in claiming that some of these facts and properties are also referents of moral terms). Its ontological commitments form a proper subset of those of non-naturalist or non-reductive naturalist accounts (excluding just their unproven *sui generis* moral facts and properties or [*sui generis*] irreducible natural ones), and it is therefore more ontologically parsimonious than those theories.

Conservatism: I argued that Goal Theory is, in the real world, conservative too, in likely preserving many of our existing core moral beliefs (as exemplified by our stock moral truisms). As I explained in sections 2.5 and 5.4, for contingent (e.g. emotional, social, and legal) reasons, I think that our true strongest desire(s) will likely be best served by the kinds of cooperative and altruistic behaviour that our strongly held and reflective moral judgements would tend to endorse (acting non-selfishly, honestly, and compassionately, for example), and vice versa. In other words, when it is properly understood and applied, I think that Goal Theory will not generally imply moral propositions that would be widely seen as being false (e.g. lying, stealing, committing murder etc.)

Motivational internalism: as I said in section 1.1, moral realism generally struggles with this criterion, because, on realism, moral judgements express beliefs, which do not seem to be intrinsically motivating. However, remember that, on Goal Theory, for *A* to sincerely judge that they ought to do *x* in *C* is for them to have a means-end belief that doing *x* in *C* would best serve the strongest desire that a fully rational and sufficiently informed version of themselves would have. However, as I explained in section 2.4, it is then hard to conceive of how it could then not plausibly follow that *A* would be *motivated* to some extent to comply with this judgement. I argued that Goal Theory may plausibly be thought of as a weak *internalist* theory, on which there is a *necessary* connection between moral judgement and motivation (for fully enlightened agents; and, assuming Sinhababu's account of the belief-desire process, for non-ideal agents too). However, even if we instead conceive of it as an *externalist* account, with only a *contingent* connection between moral judgement and motivation, we still have good reason to suppose that almost everyone (of normal psychology) who sincerely makes a moral judgement will be motivated to some extent to comply with it. Thus, however we position it, Goal Theory seems able to explain why it is that (almost) anyone who makes a sincere moral judgement would be motivated to some extent to comply with it.

Providing reasons: as I said in section 1.1, instrumentalism about reasons has difficulties explaining this, since the reason-giving power of moral requirements is then contingent upon these requirements serving one's commitments. However, on Goal Theory, the *moral* action in such and such circumstances for an agent is the one that best serves the agent's true strongest desire in those circumstances. Thus, on a Humean account of reasons, where an agent has *pro tanto* normative reason for an action just in case that action would serve some of the agent's desires, agents will have excellent reasons for compliance.[2] I discussed this is much more depth in sections 2.6 and 3.1.

[2] This connection is even more obvious on the HTR* (i.e. the variant of the Humean theory of reasons that I endorse), where an agent has *pro tanto* normative reason for an action just in case that action would serve some of the desires of a fully rational and sufficiently informed version of the agent.

The only points that I have so far not covered off are the last two, viz. moral disagreement, and semantics of moral discourse. As I said in section 1.1, objective moral realism is threatened by the former (because if morality were objective, then by all accounts we would expect to see far less moral disagreement than we do), and ethical naturalism by the latter (since it attempts to define moral properties in natural terms). Accordingly, in the remainder of this chapter I shall assess these two theoretical adequacy criteria, with the intention of motivating the conclusion that Goal Theory satisfies them too.

However, before I do that, let me first recap some findings from the thesis regarding Goal Theory's adequacy as a *first*-order moral theory. In order to assess Goal Theory's adequacy in this regard, we might, in addition to considering Argument 1 from section 2.1, want to evaluate such criteria as whether it explains what is right and wrong, giving us a clear way of getting answers to our questions about actual moral situations; whether it is comprehensive, in giving us answers, or at least a way of establishing such answers, that we can imagine applying to any situation; whether it is consistent, in not yielding conflicting results in different circumstances; when properly understood and applied, whether it yields intuitively acceptable results in almost all real-world circumstances; defuse or explain possible conflicts between self-interest and morality; and explain why we should be moral. I think that Goal Theory does indeed meet these criteria.

In summary, Goal Theory explains that what is right for a person in such and such actual moral circumstances is that which will best serve their true strongest desire in those circumstances. Whilst establishing this in practice might be a non-trivial undertaking, it is in principle clear how to go about getting the desired answers (as discussed in section 2.3). Moreover, in principle this way of establishing such answers can be applied in any moral situation in which an agent finds him or herself. In terms of consistency, because moral facts on Goal Theory supervene upon natural ones, then it will only generate different results if the relevant non-moral facts are different. Thus, if Goal Theory was to demand of *P* that they do *x* in circumstances *C,* but demand that they do ~*x* in circumstances *C**, or that some other person, *P**, does ~*x* in circumstances *C*, then then any apparent conflict would dissolve once we note that this is due to a difference in the relevant non-moral facts. Finally, In terms of yielding intuitively acceptable results in most real-world circumstances, I have discussed in some detail (especially in section 5.4) why I think it is that, when correctly understood and applied, Goal Theory achieves this. Accordingly, I think it will be found that acting in accord with such moral beliefs (e.g. non-selfishly, honestly, and compassionately), will generally better serve our true strongest desires than would acting otherwise. Finally, as explained in the previous chapter, due to its nature as an ethical egoist theory, Goal Theory avoids any possible conflict between self-interest and morality (since to act morally is always to act in one's self-interest), and supplies a ready answer to why we should be moral (because, on egoism, morality always best serves one's self-interest).

6.1 It Accounts for Moral Disagreement

As stated in section 1.1, it is generally accepted that a plausible metaethical theory should be able to account for the breadth and depth of moral disagreement that we find in the world. Such moral disagreement is widely supposed to threaten objective moral realism (and, with it, any realist theories, including Goal Theory).[3] As others have noted, there are a number of distinct (though often related) arguments from moral disagreement, and these are sometimes conflated and equivocated between (which may bolster the popularity and apparent plausibility of the claim that disagreement counts against moral realism).[4] When carefully distinguished from each other, some of these arguments carry little or no weight — begging the question against the moral realist, missing their intended target, or being easily disarmed. From those remaining, I have selected what I take to be the strongest or most influential variants, and will examine each in turn, seeking to motivate the conclusion that Goal theory resists the challenge that they pose.

Before I proceed, I should clarify that I am interested here in defending a particular variant of objective moral realism, rather than objective moral realism in general. I shall call the variant in question ‘Goal Theory Realism’. In accord with my conception of Goal Theory, I shall understand this as the view that there are moral facts and properties that are objective (in the sense of not depending upon people’s beliefs or attitudes) and reductively identical to certain natural facts and properties (of idealised human desire and cause and effect), and that there are no moral facts and properties that are not of this nature. As such, this is a strong variant of realism, and thus one that is a legitimate target for disagreement arguments.

[3] Christopher Gowans surveys this here: Christopher Gowans, 'Introduction', in *Moral Disagreements: Classic and Contemporary Readings*, ed. by Christopher Gowans (2000), pp. 1-43.

[4] E.g. Folke Tersman, *Moral Disagreement* (Cambridge: Cambridge University Press, 2006), p. xiii. David Enoch attempts to disentangle these distinct arguments: David Enoch, 'How Is Moral Disagreement a Problem for Realism?', *The Journal of Ethics*, 13 (2009), 15-50.

6.2 The IBE Version of the Argument from Moral Disagreement

John Mackie's argument from relativity is perhaps the *locus classicus* for the view that moral disagreement undermines metaethical realism and objectivity.[5] It adopts the view that moral disagreement does not deductively entail anti-realism, but instead issues an explanatory challenge to moral realists. Mackie argues that there is deep and wide-ranging disagreement in ethical matters, and that the best explanation for this is that moral judgements are not objectively true. According to him, the breadth and depth of moral disagreement is indicative of the fact that our moral judgements are merely expressions of our social commitments, with there being no need to posit some common moral reality to which agents have differential access.[6] Mackie thinks that those agents who disagree about some moral judgement intend to state the truth, but because there are no moral facts, all parties are in error.[7] As Mackie says of his argument:

> The argument from relativity has as its premiss the well-known variation in moral codes from one society to another and from one period to another, and also the differences in moral beliefs between different groups and classes within a complex community. Such variation is in itself merely a truth of descriptive morality, a fact of anthropology which entails neither first order nor second order ethical views. Yet it may indirectly support second order subjectivism: radical differences between first

[5] Mackie, *Ethics: Inventing Right and Wrong*, pp. 36-38.

[6] The emotivist Charles Stevenson agrees with this assessment: Charles L. Stevenson, 'The Nature of Ethical Disagreement', in *Exploring Philosophy: An Introductory Anthology*, ed. by Steven M. Cahn (Oxford: Oxford University Press, 2009 [1963]). By contrast, A.J. Ayer believes that there are no genuine moral disagreements, with the relevant parties really disagreeing over the relevant non-moral facts: Ayer, *Language, Truth and Logic*, pp. 102-14.

[7] Whilst Stevenson thinks that the disagreements in question are disagreements in *attitude*, involving expressions of conflicting emotions, rather than conflicting beliefs about the nature of an objective moral reality.

> order moral judgements make it difficult to treat those judgements as apprehensions of objective truths.[8]

Rather than address Mackie's original argument, I think it is more expedient for me to reformulate his argument in a stronger form, and target it directly at Goal Theory Realism. Apart from anything else, on a plausible interpretation, the original argument is consistent with Goal Theory Realism, since Goal Theory Realism is consistent with an actual universal moral error, where our moral beliefs typically reflect expressions of our social commitments, rather than reflecting an objective, independent moral reality (meaning that there may still be an objective, independent moral reality, even if moral opinions do not reflect this).[9] This does not entail the epistemologically troubling conclusion that moral facts are then radically inaccessible. Rather, I have argued (in section 2.3) that they *are* accessible, but that we have largely been looking for them in the wrong places.

Accordingly, what follows is a reformulation that I think remedies the above-mentioned issues:

[8] Mackie, *Ethics: Inventing Right and Wrong*, p. 36.

[9] For such an interpretation, see: Enoch, 'How Is Moral Disagreement a Problem for Realism?', p. 21. Also: Gowans, 'Introduction', p. 4; David Wong, 'On Moral Realism without Foundation', *The Southern Journal of Philosophy*, (1986), 95-113.

Argument 17

P1)	We observe deep and widespread moral disagreement across and within societies and periods.
P2)	This moral disagreement might be explained by agents failing to correctly apprehend an objective moral reality (due to, e.g. disagreement over the relevant non-moral facts, partiality, irrationality, and disagreement over background theory).
P3)	This moral disagreement might be explained by moral judgements being nothing other than expressions of agents' social commitments.
P4)	The latter explanation is better than the former, because, e.g. it requires us to postulate nothing over and above what we already accept.
C1)	Therefore, by an IBE, we should prefer the latter explanation to the former.
P5)	Goal Theory is committed to an objective moral reality (as defined by Goal Theory Realism).
C2)	Therefore, from C1, we have good reason to reject Goal Theory Realism.

Mackie acknowledges that nihilism is not entailed merely by the presence of widespread disagreement. After all, we find disagreement in the natural and social sciences too, and yet we do not infer from this that there are no objective facts in those realms. However, Mackie thinks that the scope of the disagreement is much greater in ethics than in these other disciplines, and that the nature of the disagreement is much more fundamental (being, for example, in principle resolvable in the natural and social sciences, but not necessarily so in ethics).

I am prepared to grant premise P1, but deny that Argument 17 undermines Goal Theory, because I deny the claim in premise P4 that we should prefer the explanation in premise P3 to that in P2 (quite the contrary in fact). As Enoch frames the possible ways of rejecting an IBE, I am then taking the route of adducing an alternative explanation for the relevant phenomenon (i.e. moral disagreement), rather than denying the need to explain the phenomenon, or denying its existence.[10] To understand why P4 is false, I think we must actually *do* the IBE, rather than merely gesturing to it.

To begin with, the question of what criteria there are for judging one explanation better than another is much debated, but popular candidates include an explanation's plausibility, simplicity, and explanatory scope and power. The first of these criteria assesses whether an explanation is consistent with what we already know to be true, where this would include such things as whether the explanation builds upon established precedents and known facts, and does not contradict other facts, for example. The second assesses an explanation's (ontological) parsimony, in terms of whether or not it introduces additional elements that not supported by independent evidence. The last two evaluate whether the explanation explains much of the evidence that we observe and makes that evidence very likely to obtain.

To introduce more clarity and rigour into the IBE, I shall employ Bayes' Theorem.[11] On this theorem, an explanation being the best explanation would equate

[10] Enoch, 'How Is Moral Disagreement a Problem for Realism?', p. 22.

[11] For defences of my implied position that IBE is compatible with Bayesian updating, insofar as the Bayesian takes into account explanatory considerations because they either do or should make use of such considerations in assigning probabilities, see: P. Lipton, *Inference to the Best Explanation* (London: Routledge, 2004); S. Okasha, 'Van Fraassen's Critique of Inference to the Best Explanation', *Studies in History and Philosophy of Science,* (2000), 691–710; J. Weisberg, 'Locating Ibe in the Bayesian Framework', *Synthese,* (2009), 125–44.

to it having the highest relative epistemic probability. As a reminder, here is the long form of Bayes' Theorem:

$$P(h|e.b) = \frac{P(h|b) \times P(e|h.b)}{[P(h|b) \times P(e|h.b)] + [P(\sim h|b) \times P(e|\sim h.b)]} \qquad [\mathbf{1}]$$

The prior probability of a hypothesis is generally understood as being composed of its plausibility and (ontological) parsimony, with its explanatory scope and power being reflected in the two consequent probability terms. An explanation being the better one for some relevant evidence would equate to it having the higher relative epistemic probability. In this case, let us compare the two hypotheses in question, viz.

> $\boldsymbol{h_1}$ = people fail to correctly apprehend an objective moral reality (as defined by Goal Theory Realism).
>
> $\boldsymbol{h_2}$ = moral judgements are nothing other than expressions of agents' social commitments.

Let me first compare the prior probabilities. In terms of the plausibility component, many agree that a hypothesis is more plausible to the extent that it respects our widely held beliefs that are resistant to alteration after reflection.[12] In this case, h_1, but not h_2, respects a particular pre-theoretical belief that is widely held and strongly resistant to alteration after reflection, viz. that our moral discourse represents reality (e.g., that genocide really *is* wrong, and charity really *is* good). So, on this view of plausibility, the hypothesis that there is no such moral reality (and thence that all moral judgements

[12] E.g. Shafer-Landau and Cuneo, *Foundations of Ethics: An Anthology*, p. 4.

are systematically and uniformly false, or are expressions of non-cognitive states, such as emotions or desires) would, *ceteris paribus*, be less plausible than the hypothesis that there is an objective moral reality (with agents engaged in moral discourse aiming to state the corresponding moral truths).

In terms of ontological parsimony, Goal Theory's reductive naturalist account claims that moral facts and properties are reductively identical to individual natural facts and properties, or combinations of natural facts and properties. As such, and contra premise P4, h_1 is committed only to entities that we already accept, in the form of facts and properties that would figure in our current natural and social science, and so it is as ontologically parsimonious as anti-realist theories, with both views positing the same natural facts and properties, differing only in that h_1 claims that some of these facts and properties are also referents of moral terms. (Of course, anti-realists might then charge the Goal Theorist with conceptual confusion, denying that any of these natural facts and properties can *be* moral facts and properties — but that is a separate issue, and one that I have already tackled in chapters 2, 3, and 4).

Thus, with h_1 *and* h_2 being equally (ontologically) parsimonious, but h_1 perhaps being more plausible than h_2 (with no good reason I am aware of to think the opposite), then I suggest it follows that the prior probability of h_1 is greater than or equal to that of h_2 i.e. $P(h_1|b) \geq P(h_2|b)$.

Let me turn now to the consequent probabilities. Taking h_2 first, if there really is no objective moral reality, and moral judgements are nothing other than expressions of agents' social commitments, then, because these commitments will likely vary widely amongst moral agents, the breadth and depth of moral disagreement that we observe in the world is as we would expect. Just as we find widespread and intractable differences in matters of taste, where there is also no objective truth of the matter, then we would expect to observe the same if there is no objective moral truth to settle moral disagreements. Hence, I think we can reasonably say that $P(e|h_2.b) \approx 1$.

But what about h_1? If there is an objective moral reality (as defined by Goal Theory Realism), would we expect to observe the breadth and depth of moral disagreement that we do observe? I suggest we would, for at least the following reasons:

- Agents are collectively basing their moral judgements upon often extensionally divergent background moral theories (and so they disagree not only on specific moral judgements, but also on the underlying moral principles upon which these judgements depend). For example, some agents will be utilitarians, others Kantians, still others virtue theorists, and so on. Many others will endorse some set of religious teachings. Yet these moral theories will sometimes yield collectively conflicting judgements in the same circumstances.[13] Of course, many will not consciously endorse and employ any of these particular moral theories, but will instead use 'commonsense' morality as their background 'theory' (where I use the term loosely — being a ragbag collection of moral imperatives derived from individuals' moral intuitions attenuated by social and cultural factors — and take this to be the default position in the absence of any of the others). In that case, as moral intuitions are plausibly evolutionary adaptations attenuated by contingent cultural factors (discussed in section 3.4), then the contingencies (along with agents' variable dispositions) imply that

[13] As Norman Daniels says: '[I]f ... moral disagreements can be traced to disagreements about [background] theory, greater moral agreement may result.' Norman Daniels, 'Wide Reflective Equilibrium and Theory Acceptance in Ethics', *Journal of Philosophy,* 76 (1979), 256-82 (p. 262).

pervasive moral disagreement will ensue.[14] In light of the foregoing, I think we would expect to see significant moral disagreement, even if there are certain near universals (e.g. that people ought not to kill innocents).

- As pointed out by David Brink and others, moral judgements generally depend upon relevant non-moral facts; yet knowledge of these facts is often difficult to obtain (or people do not make sufficient effort to do so).[15] (In Goal Theory's case, these non-moral facts include what agents' true strongest desires are, and which actions will best serve these desires in such and such circumstances.) At the same time, agents often neglect to attenuate their degree of moral certainty accordingly (or else are unaware of their ignorance). As such, even if agents agree about their background moral theory, we would still expect to find much moral disagreement, since agents will routinely disagree over the relevant non-moral facts upon which their moral judgements are based.[16] Examples of disagreements over relevant non-moral facts include whether capital punishment deters murder, whether global warming is caused by human activity, whether vaccinations cause autism, whether nuclear waste can be disposed of safely, whether genetically modified foods are safe, whether social welfare programmes help or harm economic growth, how many illegal immigrants are criminals, and whether or not torture works.

[14] See, for example: Jonathan Haidt and Craig Joseph, 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues', *Daedalus,* 133 (2004), 55–66.

[15] David Brink, *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press, 1989), pp. 198-209. Also, Boyd says that: 'careful philosophical examination will reveal ... that agreement on nonmoral issues would eliminate almost all disagreement about the sorts of moral issues which arise in ordinary moral practice.' Boyd, 'How to Be a Moral Realist', p. 213.

[16] Enoch describes such cases as not being 'moral' disagreements at all, and so would explicitly exclude them from the moral disagreement described in premise P1 of Argument 17: Enoch, 'How Is Moral Disagreement a Problem for Realism?', pp. 23-24. However, nothing turns upon this.

- Next, agents are susceptible to many kinds of irrationality (understood broadly as failures of logic or reason). Some errors of reason are sufficiently well known to have been identified as particular (formal or informal) fallacies, although many others are possible. Falling within the realm of irrationality, agents are also prone to various cognitive biases. So, even when agents are in agreement about their background moral theory, and also agree about the relevant non-moral facts upon which their moral judgements are based, we would still expect to witness moral disagreement caused by one or more parties falling prey to some form of irrationality, leading them to make faulty moral judgements.[17]
- Finally, matters of partiality (e.g. their personal prejudices, their differing self-interests, and their differential degrees of sympathy for others) routinely affect people's moral judgements.[18] Therefore, even if agents endorse the same background moral theory, agree upon all of the relevant non-moral facts, and avoid any slide into irrationality, I think we would still expect to observe significant moral disagreement.

I would argue that, taken cumulatively, the effect of the above-mentioned would be to generate deep and widespread moral disagreement, making this precisely what we should expect to observe on h_1 and these additional facts.[19]

[17] Shafer-Landau says moral realists may say that 'disagreement suggests a fault of at least one of the interlocutors [such as] ... some irrational emotional response that stands as a barrier to moral convergence.' Shafer-Landau, *Moral Realism: A Defense*, p. 218.

[18] Nicholas Sturgeon, 'Moral Explanations', in *Essays on Moral Realism*, ed. by G. Sayre-McCord (Ithaca, NY: Cornell University Press, 1988), pp. 229-56 pp. 229-30). Enoch draws special attention to the matter of agents' self-interests affecting their moral judgements: Enoch, 'How Is Moral Disagreement a Problem for Realism?', pp. 25-27.

[19] Citing Andrei Marmor, Enoch notes a curious asymmetry whereby error-theoretic proponents of the IBE argument are reluctant to attribute *moral* errors on so many matters to so many thinkers, yet are happy to attribute wide-ranging *metaethical* errors to many thinkers: Enoch, 'How Is Moral Disagreement a Problem for Realism?', p. 25.

Moreover, this explanation also accounts for the difference in the scope and nature of moral disagreements as compared to those in science, for example. Due to their provenance as evolutionary adaptations attenuated by (amongst other things) cultural (e.g. religious) factors, our moral intuitions (concerning and deriving from our concepts of fairness, equality, loyalty, authority, and sanctity, for example) are, I would argue, generally much more deeply and strongly embedded in our psyche than our scientific ones are (concerning the truth of some theoretical conjecture or other). Moreover, since moral intuitions feature so centrally in not just commonsense moral judgements, but also in the moral judgements made by philosophers, then it generally feels that much more is at stake to us in our moral judgements than in our scientific ones. As such, moral judgements are typically expressed far more fervently than scientific ones, and are much more resistant to alteration. Furthermore, moral discourse is generally more susceptible to concerns of self-interest and personal prejudice than is scientific discourse, with psychological payoffs sometimes associated with holding certain false moral beliefs — further amplifying the sense that so much more is at stake with our moral judgements than with our scientific ones.[20] In light of the foregoing, one would expect the scope of moral disagreement to far exceed that found in natural science, and for this disagreement to be more intractable — precisely what we observe.

Accordingly, I would argue that, on h_1, we would also expect to observe the sort of widespread moral disagreement that we do observe. Thus, I suggest we can reasonably say that $P(e|h_1.b) \approx 1$. And, with $P(e|h_2.b) \approx 1$, then, for each h, we

[20] Enoch also points this out: Enoch, 'How Is Moral Disagreement a Problem for Realism?', pp. 26-27. See also: Thomas Nagel, *The View from Nowhere* (Oxford: Oxford University Press, 1986), p. 148. And: Shafer-Landau, *Moral Realism: A Defense*, p. 219.

can say that $P(e|h.b) \approx 1$.[21] Moreover, since each of these hypotheses is assumed to explain the evidence, then, for each h, $P(e|{\sim}h.b) \approx 1$ (because, for example, if $h = h_1$, then $P(e|h_2.b) \approx 1$, where h_2 is part of ~h). Substituting these values into equation [1] from earlier, we find that:

$$P(h|e.b) \approx \frac{P(h|b) \times 1}{[P(h|b) \times 1] + [P({\sim}h|b) \times 1]}$$

And noting that $P({\sim}h|b)$ is the converse of $P(h|b)$, we have:

$$P(h|e.b) \approx \frac{P(h|b)}{P(h|b) + [1 - P(h|b)]}$$

That is,

$$P(h|e.b) \approx P(h|b) \quad [2]$$

So, for each h, $P(h|e.b)$ will effectively be a function of $P(h|b)$ alone. Recalling that $P(h_1|b) \geq P(h_2|b)$, then we now have:

$$P(h_1|e.b) \geq P(h_2|e.b) \quad [3]$$

Thus, when we actually carry out the IBE, then instead of finding that h_2 is a better explanation than h_1 for the widespread moral disagreement that we observe, I suggest

[21] In practice I think it is indeterminate which (if either) of h_1 or h_2 *better* fits the observational evidence overall. Moreover, I think that any difference in evidential fit is likely to be marginal. Accordingly, I shall take $P(e|h_1.b)$ and $P(e|h_2.b)$ to be equal.

that we actually find that h_1 is at least as good an explanation for this. Thus, I deny premise P4 of Argument 17, and thence conclusions C1 and C2.[22]

6.3 Irresolvable moral disagreement amongst enlightened agents

Although Mackie's IBE argument from moral disagreement seems to miss its target with my account, there are other variants of the argument from moral disagreement that may be deployed against the objective moral realist. For example, it might be argued that even if everyone were fully rational and sufficiently informed, there would *still* be moral disagreement under the same relevant conditions (I shall call this *fundamental* moral disagreement, as opposed to the *superficial* moral disagreement that obtains only in the non-ideal conditions where agents suffer from some cognitive shortcoming or other). However, we suppose that if there is objective truth in ethical matters, then those who are fully rational and sufficiently informed must be able to obtain it. Therefore, if they cannot, then it follows that (at least in such cases) there is no objective moral truth.[23] More formally:

[22] Enoch considers and rejects a variant of this argument, on which it is not deep and widespread moral disagreement in and of itself that is better explained by denying moral realism, but instead the observation that we find there is no method for resolving such disagreements (unlike in science, for example). Enoch, 'How Is Moral Disagreement a Problem for Realism?', pp. 34-39. I concur with Enoch's assessment, but will set this aside.

[23] Those who have focussed on arguments of this form include: D. Loeb, 'Moral Realism and the Argument from Disagreement', *Philosophical Studies,* (1998), 281-303; W. Tolhurst, 'The Argument from Moral Disagreement', *Ethics,* (1987), 610-21.

Argument 18

P1)	If there are possible cases of moral disagreement amongst agents in the same conditions who are fully rational and sufficiently informed, but it is the case that if there is objective truth in ethical matters then those who are fully rational and sufficiently informed must be able to obtain it, then, at least in such cases, there is no objective moral truth.
P2)	There are possible cases of moral disagreement amongst agents in the same relevant conditions who are fully rational and sufficiently informed.[24]
P3)	If there is objective truth about some matter, then those who are fully rational and sufficiently informed must be able to obtain it.
C1)	Therefore, at least in such cases, there is no objective moral truth.
P4)	Goal Theory is committed to an objective moral reality (as defined by Goal Theory Realism).
C2)	Therefore, we have reason to reject Goal Theory.

As a defender of a form of objective moral realism, I might respond to this argument in two ways. Firstly, I might deny premise P2 — arguing that genuine moral disagreement is impossible amongst fully rational and sufficiently informed agents in the same conditions. Secondly, I might concede that such disagreement is in principle possible, but then deny that objective moral realism would be undermined by the

[24] By the ‘the same relevant conditions’ here, I have in mind the same relevant biological and environmental conditions. This means that the agents’ (true strongest) desires would be the same, and the same actions would best serve those (true strongest) desires. I shall understand ‘possible cases’ here as being possible in the *actual* world, rather than in some other (nearby) possible world.

existence of genuine moral disagreement amongst fully rational and sufficiently informed agents.

The first position corresponds to what we might call a *convergentist* view. Such a view is committed to the following two claims: (1) the philosophical observation that moral realism would be undermined by the existence of moral disagreement amongst fully rational and sufficiently informed agents in the same relevant conditions; and (2) the empirical conjecture that such disagreement will not obtain amongst such agents. By contrast, the second position corresponds to a *divergentist* view, which denies that moral realism would be undermined by the existence of moral disagreement amongst fully rational and sufficiently informed agents in the same relevant conditions, and allows that such disagreement may obtain.[25]

Here I shall defend a form of convergentist view, endorsing premise P3 (at least within the sub-domain of ethical enquiry), but denying premise P2. To this end, I shall follow others (Boyd, Brink, and Sturgeon, for example) in denying the existence of any *fundamental* moral disagreement. Instead, I shall argue that any *prima facie* fundamental moral disagreement will, upon suitable inspection, turn out to be *superficial* moral disagreement instead. In that case, either 'defusing explanations' will explain away the disagreement in terms of some cognitive shortcoming or other, or else the disagreeing parties will turn out to be in relevantly different conditions. In both cases, P2 would be false.

What do I have in mind when I refer here to 'defusing explanations'? Well, I am thinking of the kind of explanations for moral disagreement that I adduced earlier in the context of the IBE argument from moral disagreement, viz. disagreement over

[25] See, for example: John M. Doris and Alexandra Plakias, 'How to Argue About Disagreement: Evaluative Diversity and Moral Realism', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity,* ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2007).

the background moral theory, disagreement over the relevant non-moral facts, irrationality, and partiality. The proponents of the argument from irresolvable moral disagreement might concede that such things account for *some* (or even most) moral disagreement, but they would then claim that when we have abstracted away all those kinds of causes of moral disagreement, some moral disagreement would remain. However, in order to raise that claim above the level of mere speculation, and actually establish the (probable) existence of fundamental moral disagreement, the anti-realist must adduce some putative example of a fundamental moral disagreement, and establish that none of the aforementioned kinds of defusing explanations (likely) applies. In the absence of that, his or her argument has no real force for anyone not already committed to the denial of moral realism — likely being question-begging instead. As Enoch argues, in a moral disagreement between two parties, neither of whom is obviously suffering from some cognitive shortcoming, the realist would simply see this disagreement as evidence of some as yet unestablished cognitive shortcoming in one or both of them — just as we would do in a case where two people disagree over the result of some arithmetical calculation, or over some visual perception or other (not positing an anti-realism in either of these cases). Unless we are already committed to an appropriate anti-realist view, the disagreement in question has no force for us.[26]

Have any plausible examples of fundamental moral disagreement been adduced by anti-realists? Here I shall critically evaluate two representative cases from Doris and Plakias. Firstly, consider the following:

> In a laboratory study (Nisbett & Cohen, 1996, pp. 45-48) subjects — white males from both Northern and Southern states attending the University of Michigan — were

[26] Enoch, 'How Is Moral Disagreement a Problem for Realism?', pp. 42-44.

> told that saliva samples would be collected to measure blood sugar as they performed various tasks. After an initial sample was collected, each unsuspecting subject walked down a narrow corridor where he was bumped by an experimental confederate who called him an ‘asshole.’ A few minutes after the incident, saliva samples were collected and analyzed to determine the level of cortisol — a hormone associated with high levels of stress, anxiety, and arousal — and testosterone — a hormone associated with aggression and dominance behavior. Southern subjects showed dramatic increases in cortisol and testosterone levels, while Northerners exhibited much smaller changes.
>
> [This study suggests] that Southerners respond more strongly to insult than do Northerners and take a more sympathetic view of others who do so. We think that the data assembled by Nisbett and colleagues make a persuasive case that a culture of honor persists in the American South. Apparently, this culture affects people's judgments, attitudes, emotion, and behaviour — down, as the hormone study might have us saying — ‘to the bone.’ In short, it seems that a culture of honor is deeply entrenched in contemporary Southern culture, despite the fact that many of the material and economic conditions giving rise to it no longer widely obtain. We are therefore inclined to postulate the existence of a fundamental disagreement between (many) Northerners and (many) Southerners regarding the permissibility of interpersonal violence.[27]

Doris and Plakias consider the diffusing explanations that I stated earlier, in an attempt to determine if any of them might explain away this disagreement, but reject them all. I agree with them that the general moral disagreement between Northerners and Southerners is not plausibly explained away as a disagreement over the relevant non-moral facts, or by irrationality or partiality. However, I then disagree with their (albeit much more tentative) rejection of a difference in background moral theory being a possible diffusing explanation in this case. Let me explain why.

I concur with Nisbett & Cohen that the evidence supports a hypothesis that a culture of honour persists in the American South (and also find plausible their

[27] Doris and Plakias, 'How to Argue About Disagreement: Evaluative Diversity and Moral Realism'.

explanation for how this culture of honour might have arisen — with the South, unlike the North, being settled by Scots-Irish herders who hailed from areas beyond the reach of central government, and whose wealth lay in stealable physical assets, causing them, like herders the world over, to cultivate a hair trigger for violent retaliation). In my view, this may then be viewed as an almost textbook example of how the ‘commonsense’ moralities of two groups of people can come to differ for contingent environmental (e.g. cultural evolutionary) reasons. However, in that case, the difference in attitude towards honour and violence between the Northerners and Southerners is then plausibly explained as a difference in their background moral ‘theory’.

Specifically, the Southerners are making moral judgements based on a background theory with particular views upon, say, masculinity and social status; whilst the Northerners are making their moral judgements based upon a background theory with *different* views upon masculinity and social status (amongst other things). As such, it is unsurprising that their judgements concerning the permissibility of interpersonal violence would sometimes diverge. Yet, under ideal discursive conditions, where all parties are fully rational and sufficiently informed, one supposes that the Northerners and Southerners would agree upon their background moral theory. This is part of what it *is* for something to be a ‘defusing explanation’ in this context. One assumes that, under ideal discursive conditions, disagreement due to some defusing explanation would dissolve. Accordingly, if the agents were in the same relevant conditions, then their moral judgements about the permissibility of interpersonal violence would align, making their previous disagreement a *superficial* one.

Doris and Plakias acknowledge that ‘differences regarding violence may be embedded in differences in background theory’, but they then go on to say:

> …notice that situating particular moral disagreements in broader theoretical disagreements doesn't always look like a *defusing* explanation; if our disagreement with the Nazis about the merits of genocide is a function of a disagreement about the plausibility of constructing our world in terms of pan-Aryan destiny, does it look more superficial for that?[28]

Here I feel that they are equivocating upon the meaning of the word ‘superficial’. As they defined it earlier in their piece, it meant a ‘disagreement where defusing explanations apply.’ And if the Nazis are using a background moral theory in which it is morally right to construct our world in terms of a pan-Aryan destiny, but we are not, then disagreements between us about the merits of genocide may be readily explained away as disagreements over background moral theory. Thus, on their definition of this, the disagreement between the Nazis and us would plausibly be a ‘superficial’ one. However, in the above quote, they seem to be taking superficial to mean something more akin to ‘insignificant’ or ‘on the surface’. Employed in this latter sense, I agree that our disagreement with the Nazis would not be a ‘superficial’ one — but Doris and Plakias are then guilty of an equivocation.

Alternatively, imagine now that we are again in ideal discursive conditions, and that both the Northerners and the Southerners agree upon their background theory (e.g. Goal Theory). Further, imagine that, in the culture in which they find themselves, the Southerners will generally act rightly on their background theory (e.g. by best serving their true strongest desires) by permitting greater interpersonal violence, whereas, in their relevantly different culture, the Northerners will not. As such, the

[28] Doris and Plakias, 'How to Argue About Disagreement: Evaluative Diversity and Moral Realism'.

Southerners are really judging that *x* (permitting more interpersonal violence) is right for them in circumstances *C*, whilst the Northerners are judging that ~*x* (not permitting this) is right for them in circumstances *C** (where *C* is relevantly different to *C**). Yet there is no contradiction here: both judgements can be true simultaneously (e.g. on Goal Theory), because the conditions in which the respective parties find themselves are *relevantly different* (contra Premise P2). Therefore, the Northerners and Southerners would be talking past one another. As such, there is again no fundamental moral disagreement.

In light of the foregoing, I find myself rejecting Doris and Plakias' view that:

> Nisbett and colleagues' work represents one potent counterexample to the convergence conjecture; the evidence suggests that the North/South differences in attitudes toward violence and honor might well persist in ideal discursive conditions.[29]

On the contrary, I would argue that the strength of this conclusion far outstrips any supporting evidence adduced by Nisbett et al., and is therefore unjustified.

Doris and Plakias also reference in their piece another putative example of a fundamental disagreement. They note that preliminary research suggests an East/West difference in terms of people's attitudes towards the kind of utilitarian calculus according to which it may be morally required for the police to prosecute and punish an individual innocent scapegoat if this would prevent rioting that will lead to greater destruction of life and property.[30] Westerners appeared more likely to endorse an *individualist* approach, on which punishing the scapegoat was morally wrong; whereas

[29] Doris and Plakias, 'How to Argue About Disagreement: Evaluative Diversity and Moral Realism'.
[30] The case of 'the magistrate and the mob', discussed in: Smart, 'An Outline of a System of Utilitarian Ethics'.

Asians were more amenable to a *collectivist* view, on which this scapegoating may be morally permissible.[31]

Once again, Doris and Plakias reject defusing explanations for this in terms of disagreements over non-moral facts, irrationality, or partiality. They then acknowledge that:

> …the putative East/West disagreement here is enmeshed in large and striking differences in *background theory*; it is entirely plausible that those with a more contextualized view of the person and a more collectivist view of society would countenance a 'one for many' approach to the magistrate and mob case.[32]

However, this concession notwithstanding, they then reject such differences in background theory as being a *defusing* explanation for the disagreement, arguing that the convergentist faces a dilemma: either ideal conditions require that all parties will agree upon the background theory, in which case it is unclear to what extent the disagreement can be understood as being between *different* cultures; or else ideal conditions do not require agreement upon background theory, in which case there is relatively little reason to expect agreement lower down, at the level of particular cases.

In response, imagine, *ex hypothesi*, that the parties are in ideal discursive conditions, and that they agree upon their background theory (Goal Theory, for example). In that case, to what extent could the disagreement in question be understood as being between *different* cultures? Here I think that the disagreement would not be a *genuine* disagreement at all, with differences in culture being an essential element of this. Specifically, it may be that relevant cultural differences (in terms of individualism versus collectivism) imply that what Asians morally ought to

[31] E.g. R. E. Nisbett, *The Geography of Thought: How Asians and Westerners Think Differently...And Why* (New York: Free Press, 2003).

[32] Doris and Plakias, 'How to Argue About Disagreement: Evaluative Diversity and Moral Realism'.

do in a ‘magistrate and the mob’ scenario is generally different to what Westerners ought to do in that scenario, because, with regard to the scenario in question, what will best serve Asians’ true strongest desires in the more collectivist culture in which they live is generally different to what will best serve Westerners’ true strongest desires in their more individualist culture. As such, the Asians are really judging on their background theory that *x* (scapegoating) is moral for them in circumstances *C*, whilst the Westerners are judging on the same background theory that ~*x* (not scapegoating) is moral for them in *relevantly different* circumstances *C**. However, once again, there is no contradiction here, since both can be true simultaneously (on Goal Theory), and thus there is no *genuine* (and thence fundamental) moral disagreement at all. Rather, the Asians and Westerners are talking past one another.

Thus, if we assume that ideal conditions require that all parties will agree upon the background theory, then it seems to me that any *prima facie* disagreement between the Asians and Westerners can be understood to a significant extent as being between *different* cultures. As such, I think that we may successfully take the first horn in Doris and Plakias’s dilemma.

In conclusion, I would argue that premise P2 of Argument 18 (i.e. there are possible cases of moral disagreement amongst agents in the same relevant conditions who are fully rational and sufficiently informed) is highly speculative, and that the putative examples of fundamental disagreement adduced have been plausibly dissolved, either by reference to defusing explanations, or by noting the relevant differences in the conditions under which the disagreeing parties find themselves. If anti-realists think that they have more compelling examples of fundamental disagreement, then I invite them to submit such examples for evaluation. In the meantime, I think that an unsupported premise P2 simply begs the question against the moral realist, meaning that Argument 18 would then persuade only those already antecedently inclined towards the denial of moral realism.

At the start of this section, I explained that it is generally accepted that a metaethical theory should be able to account for the breadth and depth of moral disagreement that we find in the world (where such moral disagreement is widely supposed to threaten objective moral realism). I have now critically evaluated two of the strongest and most dominant versions of the argument from moral disagreement, concluding in each case that Goal Theory Realism probably resists the challenge.

Notice that from the point of view of the Goal Theorist, almost nobody is applying the true background moral theory, with many relying upon ‘commonsense’ morality (with all of its manifest flaws, including its susceptibility to distortion by varying environmental factors), and others collectively employing (at least somewhat) conflicting first-order moral theories that are at best only partially true. Moreover, most moral agents are ignorant of or mistaken about relevant non-moral facts, and are prone to manifold forms of irrationality and partiality. Thus, to the Goal Theorist, it is not the least bit surprising that we observe widespread and intractable moral disagreement. However, if we all applied Goal Theory as our background theory, and took care to correctly apprehend the relevant non-moral facts and to avoid falling prey to any of the numerous kinds of irrationality and partiality, then the Goal Theorist would expect to witness the same level of agreement (about particular claims, as well as methods of discovery and justification) and tractability that we find in the sciences (finding perfect agreement under the same circumstances in the limit of ideal discursive conditions).

In conclusion, I shall consider this particular criterion for theoretical adequacy to be provisionally met, and will move on to the next.

6.4 It possesses a semantics of moral discourse

It is generally agreed that any credible metaethic must provide a theory of meaning, and successfully respond to criticisms of that theory. In this section, I shall look at a puzzle in the philosophy of language and logic that is targeted at moral naturalism, and show that it probably does not defeat my particular account, even if it might defeat others.

The problem I shall consider is Terry Horgan and Mark Timmons' Moral Twin Earth thought experiment — a prominent, contemporary version of the 'open question' argument (OQA).[33] Moore's original version of the OQA is even better known. However, even if it is cogent (and many think it is not[34]), in its standard form it has nothing to say against *non-analytic* forms of naturalism, such as Goal Theory. Accordingly, for my purpose, I shall set Moore's original OQA aside.[35]

Horgan and Timmons revive Moore's OQA (which many consider had undermined analytic naturalism, leaving only non-naturalist moral realism and anti-realism), but in a form that directly targets the more recent kind of naturalistic moral realism that aims to make true synthetic property identity statements (as opposed to analytic ones), thereby avoiding Moore's OQA in its standard form. The challenge is standardly aimed at the kind of causal semantic naturalism favoured by certain Cornell Realists, such as Richard Boyd.[36] However, when later applying their argument to Jackson's analytical moral functionalism, Horgan and Timmons state that:

[33] E.g. Horgan and Timmons, 'New Wave Moral Realism Meets Moral Twin Earth'.
[34] E.g. Dancy, 'Nonnaturalism'; Nicholas Sturgeon, 'Ethical Naturalism', in *Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press, 2006), pp. 91–121.
[35] I shall also set aside other possible variants of the OQA. For example: Eric H. Gampel, 'A Defense of the Autonomy of Ethics: Why Value Is Not Like Water', *Canadian Journal of Philosophy*, 26 (1996), 191-209; David Wiggins, 'A Neglected Position?', in *Reality, Representation, and Projection*, ed. by John Haldane and Crispin Wright (Oxford: Oxford University Press, 1993), pp. 329-36.
[36] E.g. Boyd, 'How to Be a Moral Realist'.

> We do so by applying a generic thought-experimental deconstructive recipe that we have used before against other views that posit moral properties and identify them with certain natural properties, a recipe that we believe is applicable to virtually any metaphysically naturalist version of moral realism. The recipe deploys a scenario we call Moral Twin Earth.[37]

Accordingly, I shall proceed on the basis that Horgan and Timmons would claim their thought-experimental deconstructive recipe is applicable to Goal Theory, as a metaphysically naturalist version of realism (positing moral properties and identifying them with certain natural properties).

The Moral Twin Earth argument borrows from Hilary Putnam's Twin Earth thought experiments from the 1970s.[38] In those experiments, Putnam asks us to imagine a Twin Earth, where there is a clear, odourless, drinkable liquid that falls from the sky, and which behaves just like the substance that we call 'water'. The Twin-Earthlings also call this substance 'water', but unlike on Earth, where this substance has the molecular structure H_2O, on Twin Earth the substance has the molecular structure XYZ. As such, the term 'water' has a different referent on Earth and Twin-Earth. Now, imagine that a Twin-Earthling visits Earth, and goes to see Niagara Falls. An Earthling points to the falls and says 'that is water.' However, the Twin-Earthling shakes his head and says 'that is not water.' Is there then a *genuine* disagreement between the Earthling and the Twin-Earthling? Putnam suggests our intuition is that there is not, and that they are instead talking past one another. What explains this, he

[37] Terence Horgan and Mark Timmons, 'Analytic Moral Functionalism Meets Moral Twin Earth', in *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson,* ed. by Ian Ravenscroft (Oxford: Oxford University Press, 2009), pp. 221-36 (p. 221).

[38] On Putnam's original scenario, see, for example: Hilary Putnam, 'The Meaning of 'Meaning'', in *Language, Mind and Knowledge,* ed. by Keith Gunderson (Minneapolis, MN: University of Minnesota Press, 1975).

thinks, is that a referential theory of meaning is true, and so the meaning of natural kind terms such as ‘water’ is determined by the stuff that it designates at a given world. As such, the term ‘water’ in the mouths of Earthlings and Twin-Earthlings differs in meaning. Moreover, he suggests that the English term ‘water’ rigidly designates H_2O — such that it designates H_2O at our world, and, if it is used by us to designate anything at another world, it designates something at that world just in case it designates H_2O at that world.[39]

Now, the key idea behind Horgan and Timmons’ argument against naturalistic moral realism is that Putnam’s views about how natural kind terms function causes difficulties for such theories. Ethical naturalism is typically committed to two theses: (1) an ontological thesis, whereby moral facts and properties are held to be identical to or constituted by natural facts and properties; and (2) a semantic thesis, whereby moral terms function in a similar way to natural kind terms such as ‘water’ — implying that they are not definite descriptions, but instead rigid designators of natural properties of a certain type. Accordingly, when we refer to the property of *being right*, for example, then we thereby refer to a natural property *N*, with referential use of the term designating *N*, and, if it is used by us to designate anything at any other world, it designates something at that world just in case it designates *N* at that world. According to Horgan and Timmons, the ethical naturalist’s commitments now generate a problem.

To see what this putative problem is, imagine that the word ‘right’ as used by Earthlings refers to the natural property *N*, where this is the property of being such as to best serve one’s true strongest desire, fitting with some specific egoistic normative theory, T_e. Now, imagine that Moral Twin Earth is a nearby possible world that, on the

[39] For more on rigid designators, see, for example: Saul Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980).

surface, is indistinguishable from Earth, with 'twin-moral' terms that are orthographically just like our moral terms, with significant agreement between 'right' and 'twin-right', and so on. However, perhaps due to some subtle, species-wide differences in psychological temperament, the word 'right' as used by Twin-Earthlings refers to the natural property *N**, where this is the property of being such as to maximise total happiness, fitting with some specific consequentialist normative theory, T_c (I assume here that *N* and *N** are not extensionally equivalent). Now, imagine that an Earthling was to claim that some act *x* is right for a particular agent in such and such circumstances, since doing *x* would best serve the agent's true strongest desire. Further, imagine that a Twin-Earthling responded that it is not right, because doing so would not maximise total happiness. In such a case, would there be a *genuine* disagreement between the Earthling and the Twin-Earthling (where this is understood as a disagreement in moral belief and moral theory, not in meaning or reference)? Or, would they merely be talking past one another, as in Putnam's 'water' example, with the moral terms in question differing in meaning and not being intertranslatable? According to Horgan and Timmons, our intuitive judgement in this case is that the Earthling and Twin-Earthling *do* mean the same thing when they use the term 'right', and so there really *is* a genuine disagreement, not merely a disagreement in meaning. As they say:

> We submit that by far the more natural mode of description, when one considers the Moral Twin Earth scenario, is the second. Reflection on the scenario just does not generate hermeneutical pressure to construe Moral Twin Earthling uses of 'good' and 'right' as not translatable by our orthographically identical terms.[40]

[40] Horgan and Timmons, 'New Wave Moral Realism Meets Moral Twin Earth', p. 460.

Horgan and Timmons think that, in Putnam's original scenario, if 'the two groups learn that their respective uses of 'water' are causally regulated by different physical kind-properties, it would be silly for them to think they have differing views about the real nature of water.' However, in their Moral Twin Earth scenario, Horgan and Timmons claim that 'such inter-group debate would surely strike both groups not as silly but as quite appropriate, because they would regard one another as differing in moral beliefs and moral theory, not in meaning.' As such, and unlike in Putnam's scenario, where the meaning of the term 'water' is determined by what it designates at a given world, there is a 'real' nature of 'right', independent of what it designates on Earth or Twin-Earth. From this, they conclude that if, in their use of fundamental moral terms such as 'right', the Earthlings and Twin-Earthlings would refer to distinct natural properties, with the meanings not being intertranslatable, then the corresponding reference-fixing relation posited by the moral naturalist (and thence their naturalist account) is thereby faulty.

As an example, consider Boyd's causal semantic naturalism (CSN). On that account, fundamental moral terms such as 'right' rigidly designate the unique natural properties that causally regulate their use (as with Putnam's 'water' example). However, the Moral Twin Earth thought experiment applied to CSN is set up so that *distinct* natural properties regulate the use of these moral terms on Earth and Twin-Earth (such that consequentialism of some sort true on Earth but non-consequentialism true on Twin-Earth).[41] As such, the Earthlings and Twin-Earthlings do not mean the same thing when they use these terms, and so any disagreement is not genuine, but is instead a disagreement in meaning, with the parties talking past one another — as in the water example. Yet, according to Horgan and Timmons, this experiment does not

[41] T. Horgan and M. Timmons, 'Troubles for New Wave Semantics', *Philosophical papers,* (1992), 153-75 (pp. 163-64).

generate analogous intuitions to those generated by Putnam's original scenario. Instead, the intuitions supposedly go the other way, which Horgan and Timmons take to constitute strong evidence against Boyd's view.[42]

Now consider Jackson's analytical moral functionalism (AMF). Here, the meaning of fundamental moral terms on Earth is fixed by a Lewis-style conceptual analysis based upon the commonplaces of mature folk morality.[43] Suppose that on AMF there is a single mature folk morality *M* to which all Earthlings would converge under suitably ideal reflection, with this being best systematised by some specific consequentialist normative theory. Now, the Moral Twin Earth thought experiment applied to AMF is set up such that there is a *distinct* mature folk morality *M** to which all Twin-Earthlings would converge under suitably ideal reflection, with this being best systematised by some specific *non*-consequentialist normative theory. Given this, AMF predicts that Earthlings and Twin-Earthlings would be using their phonologically and orthographically identical moral terms with a different meaning and a different referent on Earth and Twin-Earth (designating distinct unique natural properties). Thus, as before, any disagreement is not genuine, but is instead a disagreement in meaning. Yet, Horgan and Timmons once again claim that this experiment does not generate analogous intuitions to those generated by Putnam's original scenario, with the intuitions supposedly go the other way, which they take to constitute strong evidence against Jackson's view.

Expressed syllogistically, Horgan and Timmons' generic thought-experimental deconstructive recipe may be represented thus:

[42] See also: Mark Timmons, *Morality without Foundations: A Defense of Ethical Contextualism* (Oxford: Oxford University Press, 1999), p. 62.
[43] Horgan and Timmons, 'Analytic Moral Functionalism Meets Moral Twin Earth'.

Argument 19

P1)	If, on moral naturalist account *M*, fundamental moral terms, such as ‘right’, refer to distinct natural properties on Earth and Moral Twin Earth, then, on *M*, Earthlings and Twin-Earthlings will not mean the same thing by their use of these terms.
P2)	On *M*, the word ‘right’ as used by Earthlings refers to the natural property *N* (fitting with some specific normative theory T_1).
P3)	On *M*, the word ‘right’ as used by Twin-Earthlings refers to the distinct natural property $N*$ (fitting with some distinct normative theory T_2). [Per the setup of the Moral Twin Earth thought experiment]
C1)	Therefore, on *M*, Earthlings and Twin-Earthlings will not mean the same thing by their use of these terms. [Hence, any apparent moral disagreements between them would not be genuine, being disagreements in meaning instead].
P4)	If competent speakers have a strong intuitive mastery of the syntactic and semantic norms governing their language, then their intuitive judgements as to whether Earthlings and Twin-Earthlings mean the same thing by their uses of certain terms would constitute important empirical evidence for or against the hypothesis that they mean the same thing by their use of these terms.
P5)	Competent speakers have a strong intuitive mastery of the syntactic and semantic norms governing their language.

C2)	Therefore, the intuitive judgements of competent speakers as to whether Earthlings and Twin-Earthlings mean the same thing by their uses of certain terms would constitute important empirical evidence for or against the hypothesis that they mean the same thing by their use of these terms.
P6)	Our intuitive judgement, as competent speakers of our language, is that Earthlings and Twin-Earthlings mean the same thing by their use of such fundamental moral terms as ‘right’. [And hence any apparent moral disagreements between them would be genuine, being disagreements in moral belief and moral theory, not in meaning]
C3)	Therefore, we have important empirical evidence for the hypothesis that Earthlings and Twin-Earthlings mean the same thing by their use of such fundamental moral terms as ‘right’.
P7)	If we have important empirical evidence for the hypothesis that Earthlings and Twin-Earthlings mean the same thing by their use of such fundamental moral terms as ‘right’, then any account that maintains they do not mean the same thing is probably false.
C4)	Therefore, *M* is probably false.

Now, we might deny some of Horgan and Timmons’ claims here — including the intuitive judgement, as competent speakers of our language, that Earthlings and Twin-Earthlings mean the same thing by their use of such fundamental moral terms as ‘good’ and ‘right’ (premise P6); and what we might call the *semantic competence argument*, according to which these intuitive judgements constitute important empirical evidence for or against the hypothesis that Earthlings and Twin-Earthlings mean the same thing by their use of these terms (premise P4).[44] I note these objections, and will return to them later.

In what follows I shall focus upon an application of the Moral Twin Earth argument to Goal Theory specifically, setting aside Boyd and Jackson's respective views. I can find no adaptation of Horgan and Timmons' argument to target a synthetic natural reductionist account like Goal Theory, so this appears to constitute a novel application of their thought-experimental deconstructive recipe.

To this end, imagine once again that we have Earth and Twin-Earth, with the surface phenomena being indistinguishable — including a moral vocabulary that works like human moral vocabulary, using terms such as 'good' and 'bad', 'right' and 'wrong' to evaluate actions, persons, institutions, and so on. As such, were an Earthling ever to visit Moral Twin Earth, then they would be strongly inclined to translate the Moral Twin Earth terms 'good', 'right', etc. as identical to their own orthographically identical English terms. However, in line with the original thought experiment, we are to imagine that there are some subtle differences between Earth and Twin Earth (perhaps in the psychologies of the respective populations), leading to distinct referents at each world for these moral terms. To see where the differences would be located in my case, consider the following from Horgan and Timmons:

> The moral naturalist tells us a story - or at least offers a sketch of a story - about how the reference of moral terms like 'good' and 'right', when used for moral evaluation,

[44] On the semantic competence argument, see: T. Horgan and G. Graham, 'In Defense of Southern Fundamentalism', *Philosophical Studies,* (1991), 107-34. With regard to questioning this argument and the intuitions in P4, see: Eric H. Gampel, 'Ethics, Reference, and Natural Kinds', *Philosophical Papers,* (1997), 147-63.

> gets fixed. This story says something about the putative reference-fixing relation *R*, for moral terms and concepts.[45]

Boyd's reference-fixing story was about what causally regulates moral terms at each world, and Jackson's was about the systemisation of mature folk morality by means of a Lewis-style conceptual analysis. As such, this was where Horgan and Timmons located the respective differences between Earth and Twin-Earth. Now, my account's analogue to these reference-fixing stories is the chain of reasoning captured in Argument 1 from section 2.1, since it is this argument that ultimately fixes the reference of rightness in terms of some natural property (and which could be adapted to do the same for other fundamental moral terms).

If we understand 'rightness' as the state of being morally correct, then, by reference to my statement (S_2) in section 2.1 (i.e. if there is a true moral system, then its system of imperatives dictates what rational persons ought most to do), I would argue that the only essential and non-derivative element of the functional role of *being right* on Earth is that it dictates for rational persons what they ought most to do. There may be other elements, such as people thinking that an act is right leaving them disposed to do it, people being resented and punished when they knowingly do not do the right action, and so on — but these are *non*-essential, insofar as there can plausibly be acts that are right, but which people are not disposed to do, and cases where people are not resented or punished when they knowingly do not act accordingly. However, I have argued that there can be no morally right action for a rational agent in such and such circumstances that does not dictate for the agent what they ought most to do in those circumstances.

[45] T. Horgan and M. Timmons, 'Copping out on Moral Twin Earth', *Synthese*, 124 (2000), 139-52 (pp. 139-40).

Due to the indistinguishable surface phenomena (including moral vocabulary) on Earth and Twin-Earth, let me make the plausible assumptions that rightness has the same functional role on Twin-Earth, and that, in their use of the term ‘right’, the Earthlings and Twin-Earthlings both intend to refer to the natural property that fulfils this functional role on their world. Given this, then, on my account, the reference for rightness will get fixed by following the chain of reasoning captured in Argument 1. The upshot of this is that, on Earth, rightness for rational person *P* in circumstances *C* will refer to the unique natural property *N*, where this is the property of being such as to best serve *P*’s true strongest desire in *C* (thereby fitting with Goal Theory’s normative account). Horgan and Timmons might be sceptical of this result, but, as part of applying their deconstructive recipe to any given version of naturalist moral realism, they would grant such a supposition for the sake of argument (subsequently arguing that my naturalist view would still be mistaken, *even if* the supposition were correct).[46]

In order to follow Horgan and Timmons’ thought-experimental deconstructive recipe (as formalised in Argument 19), we now have to imagine that Earth and Twin-Earth differ in some subtle, bare minimum way such that when Twin-Earthlings follow the chain of reasoning detailed in Argument 1, the upshot is that the reference for ‘right’ becomes some distinct natural property *N** (e.g. the property of being such as to maximise happiness, fitting with some distinct consequentialist normative theory). Accordingly, on my account, Earthlings and Twin-Earthlings would then not mean the same thing by their use of the moral term ‘right’. However, since genuine disagreement requires that one person denies precisely what the other says, then any apparent moral disagreement between them as to whether something is ‘right’ would not be genuine, being a disagreement in meaning instead. Yet Horgan and Timmons

[46] Horgan and Timmons, 'Analytic Moral Functionalism Meets Moral Twin Earth', pp. 225-26.

would presumably claim that our intuitive judgement, as competent speakers of our language, is that Earthlings and Twin-Earthlings *do* mean the same thing by their use of such fundamental moral terms as ‘right’. This, they would then argue, constitutes important empirical evidence for the hypothesis that Earthlings and Twin-Earthlings mean the same thing by their use of such fundamental moral terms as ‘right’ — thereby rendering my account probably false.

How might I respond to Horgan and Timmons’ argument? Well, with echoes of my strategy in section 5.4, I would adopt a two-pronged approach, arguing: (1) given the assumptions and stipulations of the thought experiment, it is very improbable that the upshot of the Twin-Earthlings following the chain of reasoning captured in Argument 1 will be that the reference for ‘right’ is genuinely some distinct natural property, *N**; (2) but if ever this was the case in some possible (albeit bizarre) circumstances, then I would ‘bite the bullet’, accepting that this is genuinely implied, but then denying that this outcome would be false *in those circumstances*, notwithstanding any intuition to the contrary. Let me explain.

Firstly, I would argue that my reference-fixing story (captured in Argument 1) is just not sensitive to the kinds of subtle differences that Horgan and Timmons posit between Earth and Twin-Earth in the Boyd and Jackson cases (e.g. with regard to the psychology of the Earthlings and Twin-Earthlings, where Twin-Earthlings tend to experience the sentiment of guilt more readily and more intensively, and tend to experience sympathy less readily and less intensively, than do Earthlings[47]). Having already granted that the functional role of *being right* on Twin-Earth is that it dictates for rational persons what they ought most to do, then we have now reached premise P3 of Argument 1. Thus, any difference in the upshot of this argument must manifest itself after that point.

[47] Horgan and Timmons, 'New Wave Moral Realism Meets Moral Twin Earth', p. 459.

Looking at premise P4, could it be that some subtle difference between Earth and Twin-Earth means that what a rational person ‘ought most to do’ on Earth is very probably identical to ‘what they will do, when fully rational and sufficiently informed, but that this is not so on Twin-Earth? Given my particular reductio defence of this claim in section 2.1, this seems unlikely. How about premise P5: could it be the case on Earth but not on Twin-Earth that what rational persons will do, when fully rational and sufficiently informed, is probably what they would desire most to do, when fully rational and sufficiently informed? This step in the argument is based upon the action-based theory of desire (on which for a person to *desire* Φ is for the person to be disposed to take whatever actions they believe are likely to bring about Φ). So, might this theory of desire hold on Earth, but, due to some subtle difference in psychology or suchlike, fail to hold on Twin-Earth? Again, I find this unlikely. Whatever plays the role of a desire in the Twin-Earthlings (regardless of the internal constitution that realizes this mental state), if one grants the action-based theory of desire on Earth (as Horgan and Timmons would do, as part of their deconstructive recipe where they grant my reference-fixing story), then it is difficult to conceive of some bare minimum, subtle difference that could lead to the theory being false on Twin-Earth. However, if neither premise P4 nor P5 yields some substantively different result on Twin-Earth, then the Twin-Earthlings would find that the referent of rightness on their world is the same as that on Earth.

I think the only plausible way that some difference between Earth and Twin-Earth could result in Argument 1 yielding distinct referents of ‘right’ on each world would be if this difference was of such magnitude that the two worlds would then have significantly different surface phenomena too, e.g. the two worlds having distinct functional roles for rightness. However, this would then be inconsistent with the initial assumptions and stipulations of the thought experiment (and would mean that the respective populations of the two worlds may then have different referential intentions, in which case any apparent moral disagreements would not being genuine, but disagreements in meaning instead), so I shall set it aside.

In light of the above, I would argue that if we follow Horgan and Timmons' thought experiment as it would be applied to my account (with its assumptions and stipulations about the indistinguishability of surface phenomena and suchlike), then it is far more plausible that the *same* natural property will genuinely satisfy the functional role of rightness on both worlds, but that the Twin-Earthlings have erred in their reference-fixing on their world (by making a mistake in the chain of reasoning captured in Argument 1), and so are mistakenly referring to the *wrong* natural property. (Remember that, as part of their deconstructive recipe, Horgan and Timmons would grant my reference fixing on Earth, so if one or other population is wrong, then by implication it will be the Twin-Earthlings.) For example, in their use of 'right', the Twin-Earthlings might be mistakenly referring to the distinct natural property of being such as to maximise total happiness.

Thus, in common with David Copp, I think our naturalistic semantic theory will incorporate a theory of error (as Copp thinks any coherent theory will), to allow for the possibility that agents' *beliefs* about the nature of the property that genuinely fulfils the functional role of rightness on their world are misaligned with the *actual* nature of the property concerned (noting that on Putnam's semantics, it would be the *latter*, in conjunction with the actual content of the referential intention, that determines the referent of the term in question).[48] As a result, if agents' beliefs are mistaken in this regard, then they can have the wrong referent for the moral term.

[48] D. Copp, 'Milk, Honey, and the Good Life on Moral Twin Earth', *Synthese*, 124 (2000), 113-37 (pp. 124-34).

In light of the foregoing, my account plausibly gives us sameness of reference and sameness of meaning, and so does not conflict with Horgan and Timmons' posited intuitive judgement, as competent speakers of our language, that Earthlings and Twin-Earthlings mean the same thing by their use of such fundamental moral terms as 'right'. In that case, any apparent moral disagreements between Earthlings and Twin-Earthlings about what is right would be genuine ones, being disagreements in moral belief and moral theory, not in meaning (thereby giving Horgan and Timmons what they would want). This argument may be expressed syllogistically as follows:

Argument 20

P1)	If there is no plausible means by which some subtle difference between Earth and Twin-Earth (that is consistent with the surface phenomena, including moral vocabulary, being indistinguishable on the two worlds) could result in Argument 1 yielding distinct referents (*N* and *N**) for the fundamental moral term *T*, where *T* = 'right', on each world, then it is far more plausible that the *same* natural property genuinely fulfils the functional role of rightness on both worlds, but that the Twin-Earthlings have erred in their reference-fixing on their world (by making a mistake in the chain of reasoning captured in Argument 1), and so are mistakenly referring to the *wrong* natural property (i.e. *N**).
P2)	There is no plausible means by which some subtle difference between Earth and Twin-Earth (that is consistent with the surface phenomena, including moral vocabulary, being indistinguishable on the two worlds) could result in Argument 1 yielding distinct referents (*N* and *N**) for the fundamental moral term *T* on each world.

C1)	Therefore, it is far more plausible that the *same* natural property genuinely fulfils the functional role of rightness on both worlds, but that the Twin-Earthlings have erred in their reference-fixing on their world, and so are mistakenly referring to the *wrong* natural property (i.e. *N**). [Therefore, we must incorporate a semantic error theory.]
P3)	If, in their use of *T*, the Earthlings and Twin-Earthlings both intend to refer to the natural property that fulfils the functional role of rightness on their world, if this functional role is the same on both worlds, and if the same natural property, *N*, genuinely fulfils this functional role on both worlds, then the Earthlings and Twin-Earthlings will mean the same thing by their use of *T*.
P4)	In their use of *T*, the Earthlings and Twin-Earthlings both intend to refer to the natural property that fulfils the functional role of rightness on their world, and this functional role is the same on both worlds. [From the indistinguishability of surface phenomena assumptions.]
C2)	Therefore, the Earthlings and Twin-Earthlings plausibly mean the same thing by their use of *T*. [C1, P3, P4]
P5)	If Earthlings and Twin-Earthlings plausibly mean the same thing by their use of *T*, then any apparent moral disagreements between them about what is ‘right’ would plausibly be genuine, being disagreements in moral belief and moral theory, not in meaning.
C3)	Therefore, any apparent moral disagreements between the Earthlings and Twin-Earthlings about what is ‘right’ would plausibly be genuine, being disagreements in moral belief and moral theory, not in meaning. [C2, P5]

Thus, I directly challenge Argument 19, as it would be applied to my account. Specifically, I deny premise P3 (i.e. on *M*, the word ‘right’ as used by Twin-Earthlings refers to the distinct natural property *N**). As such, my account preserves sameness of meaning and reference, contra Argument 19.

In a response to Horgan and Timmons’ argument, Copp argues that in order to resist Horgan and Timmons’ argument, the moral naturalist would need to develop a semantic account of moral terms that:

> (1) preserves the idea that the two groups featured in our argument use moral terms with the same meaning (and thereby preserves the intuition that the groups engage in genuine moral disagreement); (2) implies that both groups, in their moral uses of such terms as 'right' and 'good', are referring to the same properties of rightness and goodness; and yet (3) the moral judgments of one of the groups are mistaken (thus avoiding unwanted relativist implications).[49]

Horgan and Timmons do not deny that any account that delivered such a semantic story would resist their argument. Yet, they think that Copp’s account fails in this regard. Moreover, they think that there is good reason to be extremely pessimistic about the prospects of delivering one at all. However, I would argue that such pessimism is unwarranted, given what I have just said regarding my account. Let me explain.

On my account, the functional role of ‘rightness’ is the same on Earth and Moral Twin Earth (a very plausible assumption, given the setup of the thought experiment), and I argue that this functional role is genuinely satisfied on both worlds by the *same unique* property, *N*, where *N* = being such as to best serve persons’ true

[49] This is as summarised by Horgan and Timmons. T. Horgan and M. Timmons, 'Copping out on Moral Twin Earth', ibid., 139-52 (p. 142). For Copp’s original, see: D. Copp, 'Milk, Honey, and the Good Life on Moral Twin Earth', ibid., 113-37 (pp. 124-34).

strongest desires (given the chain of reasoning captured in Argument 1, and its insensitivity to subtle changes). As such, we have a semantic view according to which we have sameness of meaning and sameness of reference, with the two groups intending to use the moral term ‘right’ to refer to the same property on each world. Accordingly, the intuition that the groups engage in *genuine* moral disagreement is preserved. Consequently, the first two criteria from above appear to be satisfied. What is more, on my semantic error theory, the moral judgments of one of the groups is mistaken, with the Twin-Earthlings using the term ‘right’ with the *intention* to refer to the property, *N*, that genuinely satisfies the functional role of rightness on Moral Twin Earth, but mistakenly referring to the wrong property, *N** (e.g. being such as to maximise total happiness). By such means, we avoid unwanted relativist implications. Therefore, my account plausibly succeeds in delivering a semantic story that accomplishes Copp’s aforementioned three feats.

So, as things stand, I think I have at least motivated the conclusion that my moral naturalist account succeeds where Boyd and Jackson’s arguably fail, in delivering a plausible form of synthetic moral naturalism on which the corresponding Earthling and Twin-Earthling moral terms appear to express the same natural properties (at least for *rightness*, and I suggest that the same applies for other moral terms too), and therefore sameness of meaning is preserved.

How might Horgan and Timmons respond to this? Well, they pose a dilemma for any proposed version of moral naturalism:

> The first horn is that the putatively reference-fixing relation *R* might fail to fix *determinate* reference-relations between moral terms and certain natural properties, because there are too many eligible natural properties that satisfy the constraints imposed by *R*. For instance, perhaps the *R*-constraints are satisfied by a class of natural properties - functional properties, say - that collectively satisfy some

> consequentialist moral theory T_C, and yet the *R*-constraints are *also* satisfied by another class of natural properties — also functional properties, say — that collectively satisfy some deontological moral theory T_D?[50]

In terms of the first horn, ‘right’ is appropriately *R*-related to natural property *N* on my account if the functional role of rightness is genuinely fulfilled by *N* (as determined by means of Argument 1). However, as implied by my analysis in section 2.1, I think that the only essential part of the functional-role profile of rightness (viz. dictating for rational persons what they ought most to do) is *uniquely* satisfied on Earth by the natural property *N*, where *N* = being such as to best serve persons’ true strongest desires. Hence, because on my proposal there is only *one* eligible natural property that satisfies the constraints imposed by *R*, then I would argue that the first horn of the dilemma does not trouble my proposal.

With regard to the second horn of their dilemma, Horgan and Timmons say that:

> The second horn of the dilemma arises if one grants (at least for argument’s sake) that the proposed reference-fixing relation R suffices to pin down some unique class of natural properties as the putative referents of moral terms. We now tell a story about two groups, one on earth and one on Twin Earth (though the Twin Earth device really is not necessary) where the natural properties *R*-linked to the moral terms as used by both groups are different on earth than on Twin Earth. We grant our opposition (at least for the sake of argument) its assumption that there is a single natural property to which all human uses of 'right' (and its synonyms in other languages) are appropriately *R*-related. We then go on to suppose, for the sake of vividness, that on earth the term 'right', in its moral uses, is appropriately *R*-related (according to the account on offer) to the sort of natural property satisfying a consequentialist moral theory T_C (say, a functional property characterizable in terms of T_C) while on Moral Twin Earth, ‘right’ is appropriately R-related to the sort of natural property satisfying a nonconsequentialist, deontological moral theory T_D (say, a functional property

[50] T. Horgan and M. Timmons, 'Copping out on Moral Twin Earth', ibid., 139-52 (p. 140).

> characterizable in terms of T_D). Now the point of constructing this sort of scenario is that it reveals that the proposed version of naturalistic moral realism is committed to objectionable relativism: in cases where the two groups are engaged in what appears to be substantive moral disagreement, it turns out that according to the proposal, what each side says is (in their own mouths) true, because for each side the moral terms stand in the reference-determining relation to certain natural properties such that the moral statement, in the speaker's mouth, is true. But then, the groups are really talking past one another; there is no genuine disagreement in the beliefs expressed by the assertions that employ the moral and the twin-moral vocabulary. Depending on the details of the specific version of naturalistic moral realism under consideration, the view turns out either to be guilty of one or another kind of objectionable relativism. On the one hand it is guilty of chauvinistic conceptual relativism if it is committed to saying that the orthographically identical human and twin-human words ‘right’ have different meanings, as well as different referents (analogously to the human and twin-human words ‘water’ in Putnam's original Twin Earth scenario). On the other hand, if the view implies that ‘right’ has the same meaning in English as in Twin English, but has different referent properties when employed by humans and twin humans respectively, then it is guilty of standard relativism, since then the very same moral judgment may, e.g., be true for earthlings but false for twin earthlings.[51]

However, on Horgan and Timmons’ thought experiment, the possible worlds are sufficiently close that, on my proposal, it is very plausibly the case that ‘right’ is appropriately R-related to the *same* unique natural property on both Earth and Twin-Earth, viz. *N*, where *N* = being such as to best serve persons’ true strongest desires (per Argument 20). It only *appears* to be appropriately R-related to distinct natural properties on each world in its moral uses because the Twin-Earthlings are mistakenly referring to the *wrong* property (per my semantic error theory), and so the natural property to which they think ‘right’ is appropriately R-related was not their referential intention. As such, the two groups using the moral term ‘right’ *intend* to refer to the *same* property on each world.

[51] Horgan and Timmons, 'Copping out on Moral Twin Earth', p. 140.

Now, in light of the foregoing, let me see if my view is committed to objectionable relativism, as Horgan and Timmons would presumably want to claim. First, in not being committed to saying that the orthographically identical human and twin-human words ‘right’ have different meanings, as well as different referents (analogously to the human and twin-human words ‘water’ in Putnam's original Twin Earth scenario), then my version of naturalistic moral realism is not guilty of ‘chauvinistic conceptual relativism’, as Horgan and Timmons conceive of this. Second, my view does imply that ‘right’ has the same meaning in English as in Twin-English, but has different referent properties when employed by humans and twin-humans respectively. So, is it guilty of *standard relativism* (as Horgan and Timmons conceive of this), with the very same moral judgment being true for Earthlings but false for Twin-Earthlings? No, because on my proposal we only have different referents because the Twin-Earthlings are mistakenly referring to the wrong properties. As such, any appearance of standard relativism is illusory, since, in reality, the very same moral judgement would be either true or false for both the Earthlings and Twin-Earthlings, regardless of their beliefs about this.[52] Thus, I would submit that my view is not committed to objectionable relativism.

Horgan and Timmons admit, of course, that they have not tested every possible version of moral naturalism. Yet, as they say:

> It is very hard to see what sort of story about reference-fixing, for moral and twin-moral terms, could manage to break that symmetry in a way that both (1) allows for genuine disagreement (rather than being committed to objectionable relativism, given the symmetry), and (2) yields significant determinacy of moral facts and moral truths (rather than surrendering determinacy, because of the symmetry). Either the putative

[52] Though some local relativism of kind I spoke about earlier regarding Asians and Westerners is permitted, where some moral judgement may be true for one group but false for the other, because they find themselves in relevantly different local conditions.

> reference-fixing relation R will fail to secure determinate reference at all (because there are too many natural properties as eligible referents for moral terms), or else R will link moral terms to different natural properties on earth and on Moral Twin Earth - so that the claim that R fixes reference ends up committed to objectionable relativism.[53]

However, I would claim that my ‘story’ succeeds here, insofar as it *does* allow for genuine disagreement (since *R* plausibly links moral terms to the *same* natural properties on Earth and on Moral Twin Earth, notwithstanding the Twin-Earthlings being mistaken about the referent in question); and it *does* yield significant determinacy of moral facts and truths (since, in the case of ‘right’, for example, there is only *one* natural property that is an eligible referent for the moral term).

Horgan and Timmons point out a potential problem with an approach that appeals to commonalities in referential intentions of Earthlings and Twin-Earthlings. As they say, if we suppose that both groups use moral terms with the intention of picking out e.g. those properties that bear on human flourishing as judged from a standpoint of impartiality, then that notion is sufficiently vague that it might lead to indeterminacy of reference again — where different, incompatible moral theories are equally compatible with the generic notion of flourishing and impartiality. However, in contrast to this, I think that my conception of the functional role of ‘right’ (whereby it dictates for rational persons what they ought most to do) entails a unique property that fulfils the role (and likewise, I suggest, for other fundamental moral terms, such as ‘good’ and ‘wrong’).

In light of this, and as required by Horgan and Timmons, I think that I *can* legitimately ‘appeal to common referential intentions associated with moral thought and discourse in an attempt to tell a story about moral reference that yields

[53] Horgan and Timmons, 'Copping out on Moral Twin Earth', p. 141.

determinacy of reference for terms like ‘good’ and ‘right’.’[54] Unlike with vague notions like flourishing and impartiality, where the appropriate functional properties of both a consequentialist theory T_C and deontological theory T_D may be compatible, I maintain that T_C and T_D (for example) are not compatible with the referential intentions of the two groups in their use of terms such as ‘right’ (as I conceive of them). Only my account is compatible, or so I have argued, and so there is no moral indeterminacy.

In summary, and contra Horgan and Timmons, I would argue that it is very implausible that some subtle difference between Earth and Twin-Earth (that is consistent with the surface phenomena, including moral vocabulary, being indistinguishable on the two worlds) could result in Argument 1 genuinely yielding distinct referents for fundamental moral terms such as ‘right’ on each world. And, in that case, any apparent moral disagreements between the Earthlings and Twin-Earthlings about what is ‘right’ would be genuine ones, being disagreements in moral belief and moral theory, not in meaning (thereby aligning with Horgan and Timmons’ intuition about sameness of meaning).

Horgan and Timmons criticise Copp for suggesting that he has provided a way around the Moral Twin Earth argument, when they suggest he has merely adduced a wish list for what a non-analytic naturalist account would have to accomplish in order to survive their argument, without presenting an account that actually fills this wish list.[55] However, I would argue that, in the case I have been evaluating, my account *would* plausibly fill Copp’s ‘wish list’, overcoming extreme indeterminacy of moral reference and moral truth, whilst simultaneously avoiding objectionable moral relativism.

[54] Horgan and Timmons, 'Copping out on Moral Twin Earth', p. 145.
[55] Horgan and Timmons, 'Copping out on Moral Twin Earth', p. 149.

Despite what I have argued, what if there really can be some possible (albeit bizarre) scenario on my account in which subtle difference between Earth and Twin-Earth (that are consistent with the surface phenomena, including moral vocabulary, being indistinguishable on the two worlds) could genuinely yield distinct referents for fundamental moral terms such as ‘right’ on each world (with neither population mistakenly referring to the wrong properties)? In such a scenario, my account would not preserve sameness of meaning or reference, and so would conflict with Horgan and Timmons’ semantic intuition (premise P6 of Argument 19). As such, we cannot keep both — so which one should we prefer? In such a case, I would have more trust in my account than in the conflicting intuition, and so would adopt the second strategy I described earlier, viz. ‘biting the bullet’ by accepting that difference of meaning and reference is implied by my account, but then denying that this would be false *in those circumstances*.

I reach the above conclusion by weighing the epistemic merits of the two options. On the one hand, we have an account that is supported by a plausible (albeit defeasible) positive argument (i.e. Argument 1), which seems to have survived all the objections that I have critically evaluated, and which is *prima facie* theoretically adequate. On the other hand, we have a semantic intuition. How much credence should we give to this intuition? Relatively little, I would suggest. Firstly, I argued in section 3.4 that intuitions are routinely unreliable (even when they are strong and widespread, like the intuitions that the Earth is flat and stationary), and that we have no generally accepted means to distinguish trustworthy intuitions from untrustworthy ones. Thus, beliefs based (solely) on intuitions are probably not justified. Moreover, I would argue that even if semantic intuitions were generally reliable, whether there is a ‘real’ nature of ‘rightness’ — such that Earthlings and Twin-Earthlings would always mean the same thing by their use of the term, even if they have distinct referents — is not the kind of question that can be settled by reference to intuitively understood semantic rules.

Here I would agree with Janice Dowell, who argues that even if we hold that our judgments about XYZ have probative value for semantic theorizing about 'water', our judgments about meaning in hypothetical Twin-Earth thought experiments need not have probative value for constructing a semantics for our moral terms (with semantic theories being contingent and empirical, more like biological theories than like paradigmatically philosophical ones, such as theories about the nature of knowledge or of normativity).[56] I shall set aside further discussion of this, but I think I have at least shifted the burden of proof back onto those who would treat such semantic intuitions as having probative value.

So, in summary, and with reference to four possible defensive stratagems adduced by Horgan and Timmons', I would not go so far as to claim that, when applied to my account, the Moral Twin Earth thought experiment does not describe a genuinely possible scenario (in line with their first option), though I allow that it might not.[57] Rather, I combine a weaker form of this stratagem with one that they do not present. Specifically, I find it very improbable that the Moral Twin Earth thought experiment describes a genuinely possible scenario on my account. At the same time, I find it far more plausible that the Earthlings and Twin-Earthlings have the *same* referential intention when they use the moral term 'right' (i.e. to refer to the natural property that genuinely fulfils the functional role of 'rightness'), and that this natural property will be the *same* on both worlds, but that the Twin-Earthlings have erred in their reference-fixing, and so are mistakenly referring to the *wrong* natural property

[56] J. Dowell, 'The Metaethical Insignificance of Moral Twin Earth', in *Oxford Studies in Metaethics (Vol. 11)*, ed. by R. Shafer-Landau (Oxford: Oxford University Press, 2015).
[57] Horgan and Timmons, 'Troubles for New Wave Semantics', pp. 168-69.

(per my semantic error theory). As such, sameness of reference and meaning are preserved, and so any apparent disagreement between the Earthlings and Twin-Earthlings would be genuine. This outcome aligns with Horgan and Timmons' semantic intuition, and seems to avoid any objectionable relativism and indeterminacy of reference.

However, if there were any genuine possible scenario in which, when applied to my account, the Moral Twin Earth thought experiment really would entail difference of meaning and difference of reference, then, in those circumstances, I would accept this, trusting my account over any contrary semantic intuition (and thereby taking Horgan and Timmons' second stratagem). As such, in those circumstances, I would accept that any apparent disagreements are not genuine ones, but are instead disagreements in meaning. Horgan and Timmons argue that this stratagem carries with it an 'enormous' burden of proof. At the very least, they say:

> [T]he ethical naturalist who would go the avoidance route must plausibly explain (i) why peop1e's meaning-intuitions about moral terms are so strong and so widespread even though they are allegedly mistaken, and (ii) why those intuitions don't work the same way they do in Putnam's original cases.[58]

In response, I would argue firstly that it is not shown that people's 'meaning-intuitions' *do* so strongly and so pervasively line up with what Horgan and Timmons' posit. This might be so, but Horgan and Timmons merely assert it. However, if, for the sake of argument, I grant this assertion, then I would suggest that there may be good evolutionary reasons (to do with enforcing cooperation and inhibiting defection amongst groups) for a strong and widespread intuition that there is a 'real' nature of

[58] Horgan and Timmons, 'Troubles for New Wave Semantics', p. 174.

morality, that others share our perception of this nature, and that we and they cannot just interpret this nature in our own idiosyncratic ways. By contrast, I submit that there has been no evolutionary pressure for us to feel that there is a 'real' nature of water (whose chemical nature we have only very recently come to know anyway), and so we just do not feel any analogous intuition. I would suggest it is this that our intuitions are keying into, rather than any probative value inherent in tacitly understanding semantic rules.

Horgan and Timmons also say that one of the defining characteristics of a moral code is that it performs an action-guiding role for members of the community in which it is in force; this normative aspect amounts to a semantic constraint for interpreting the practices of a community as moral practices, and so is plausibly taken to be built in to the meaning of moral terms like 'good' and 'right'; and that this helps to explain why our intuitions go the way they do.[59] Notice, however, that my account respects such a semantic constraint, with the moral code on both worlds performing an action-guiding role for members of the community in which it is in force (being derived from the claim that morality's system of imperatives applies to all rational persons, governing behaviour that affects others, and should never be overridden), even if it could ever be the case that this did not entail Goal Theory.

This option still avoids chauvinistic conceptual relativism, insofar as it does not commit me to claiming that actual or possible agents who have a referent of rightness different from that of Earthlings would not possess the concept of rightness at all. Rather, I think at the level of functional role they would possess the *same* concept that we do. Moreover, at the level of the property that fulfils this role, they would still possess a concept, but it would just be a different concept to ours. It may be guilty of *standard* relativism, with the very same moral judgment being true for

[59] Horgan and Timmons, 'Troubles for New Wave Semantics', p. 170.

Earthlings but false for Twin-Earthlings. However, in such a scenario, I suggest that this really would be the correct representation of the moral circumstances.

Accordingly, I would argue that my account plausibly succeeds where Boyd and Jackson's accounts arguably fail, accomplishing something that Horgan and Timmons doubted could be done. If they deny this success, then I think the burden of proof is shifted back onto them to explain why.

6.5 Conclusions

I began this chapter by reviewing a number of considerations that are generally acknowledged to bear upon the theoretical adequacy of any metaethical theory that seeks (as Goal Theory does) to answer the basic metaphysical question: 'what is the nature of moral reality?' In particular, Shafer-Landau and Cuneo identify eight criteria that are widely thought to be necessary for such a theory to meet in order to be considered theoretically adequate. I explained why I had already effectively established Goal Theory's compliance with the first six. The remaining two were the condition that any plausible metaethical theory should account for the relatively greater depth and breadth of moral disagreement, as compared with other areas of supposed objective truth; and the condition that it has a semantics of moral discourse, supplying plausible answers to well-known semantic puzzles. Accordingly, my objective in this chapter was to establish Goal Theory's compliance with these last two conditions.

With regard to the first of these, I explained that there are a number of different arguments from moral disagreement, and that these are sometimes conflated and equivocated between. However, referencing the distinctions outlined by Enoch, I selected for evaluation two of the most influential and forceful variants of the argument, viz. the IBE version of the argument from moral disagreement (as advocated by Mackie), and the argument from irresolvable moral disagreement amongst fully enlightened agents. In both cases, I found that my account plausibly resisted the argument in question. Perhaps anti-realists can present better versions of the argument from disagreement. However, for now, I submit that my account is not undermined.

Next, I assessed the second of the above-mentioned conditions. Specifically, I critically evaluated Terry Horgan and Mark Timmons' Moral Twin Earth thought experiment as this would be applied to my account, arguing that it probably fails. Specifically, I argued first that there is no plausible means by which some subtle difference between Earth and Twin-Earth (that is consistent with the surface phenomena, including moral vocabulary, being indistinguishable on the two worlds) could result in Argument 1 yielding distinct referents for the moral term 'right' on each world; and that the Twin-Earthlings have much more likely fallen into semantic error, mistakenly referring to the wrong natural property (thereby preserving sameness of meaning and reference, and thence genuine disagreement). Yet, if ever this were to be the case in some genuinely possible scenario, then I would 'bite the bullet', accepting that this is implied, but then denying that this result (with the Earthlings and Twin-Earthlings talking past one another) would be false *in those circumstances*, notwithstanding any semantic intuition to the contrary.

In conclusion, I submit that Goal Theory plausibly satisfies *all* of the previously identified criteria for theoretical adequacy, including those with which realist and naturalist metaethical views generally struggle.

Chapter 7

Overall conclusions

At the start of this thesis, I asked what the true nature of moral reality is, what the true content of morality is, and how we might best make progress in the quest to establish these. I then observed that when we survey the contemporary metaethical and normative landscape, we find a problem, insofar as all of the familiar theories of the nature of moral reality or content of morality struggle to answer serious objections faced by the theoretical viewpoints to which they belong, and all seem to inherit the poor scores of their respective viewpoints on certain widely accepted criteria for theoretical adequacy..

After this initially pessimistic assessment, I struck a more optimistic note, suggesting that there is a novel theory that plausibly resists the serious objections faced by the theoretical viewpoint to which it belongs, and that meets all of the widely accepted criteria for theoretical adequacy. As such, we may be in a position to make useful progress in our efforts to establish the true nature of moral reality and the true content of morality (and other related matters where ethics comes into contact with metaphysics, epistemology, philosophy of mind, and philosophy of language). To that end, we might unpack the novel theory's first and second-order commitments along various dimensions, advance positive arguments for the theory and its commitments, critically evaluate and respond to dominant objections to these commitments, and assess the theory for adequacy against some applicable set of criteria. I made clear that the ideal aim would not be to add another partially inadequate theory to the landscape, but instead to adduce a theory that substantively *improves* upon existing ones, by repairing defects in these theories or suffering from fewer or less severe vulnerabilities. This is the approach that I have taken in my thesis, proposing Carrier's Goal Theory as the *prima facie* adequate novel theory.

So, what has been achieved by my research? Well, as noted in chapter 1, Goal Theory is unknown within the academic literature, having so far only received a relatively high-level treatment from Carrier. As such, it was significantly underspecified, and therefore in need of much unpacking in order to identify and critically evaluate its commitments. To that end, I began in chapter 2 by formulating an original positive argument for Goal Theory, and continued by unpacking its commitments along various dimensions of ethics and metaethics.

In terms of metaphysics, I found Goal Theory to be a realist and reductive naturalist account (with its moral facts being reductively identified with natural facts of idealised human desire and cause and effect), where there is no requirement to add unproven *sui generis* non-natural or (*sui generis*) irreducible natural entities to our ontology (as non-naturalist or non-reductive naturalist accounts do). Indeed, I found it to be as ontologically parsimonious as anti-realist accounts that deny an objective moral reality, positing the same natural facts and properties, and differing only from anti-realist theories in claiming that some of these facts and properties are also referents of moral terms. With regard to epistemology, I found that by locating the domain of morality within the familiar natural world, Goal Theory's moral facts and properties are in principle discoverable by the familiar methods of science, with no requirement to posit some special faculty or other means by which we may come to know them (as non-naturalist accounts do). As for philosophy of language, Goal

Theory is a representationalist and truth-apt account, whose moral sentences are like other statements of fact (in representing a way reality could be). In terms of moral psychology, Goal Theory is a cognitivist account, on which moral judgements express agents' beliefs. As a realist account, it also holds that some of these beliefs are *true* (unlike error theoretical accounts). Moreover, it is able to offer a plausible explanation of why it is that (almost) anyone who makes a sincere moral judgement would be motivated to some extent to comply with it. This is all attractive theoretical territory to occupy.

I then answered some common questions and challenges, and pointed the way to a more detailed evaluation of some particularly important objections to its metaethical and normative commitments, viz. (1) that there are categorical normative reasons, contra Goal Theory's Humean account; (2) that normative facts and properties are 'just too different' from natural facts and properties to be reducible or identical to them, contra Goal Theory's naturalistic account of normativity; and (3) that ethical egoism succumbs to a combination of internal and external criticisms, spelling serious trouble for Goal Theory's egoist account. During the course of chapters 3, 4, and 5, I critically evaluated these three objections, concluding in each case that my account is probably resistant. I then showed in chapter 6 how Goal Theory plausibly satisfies all eight theoretical adequacy criteria listed in section 1.1.

I would suggest that in offering a characterisation of moral facts and properties as natural ones that plausibly meets *all* of these adequacy criteria — in addition to offering plausible answers to the above-mentioned objections — my account improves upon rival ones. No account is free of all flaws, but I would argue that my account has fewer, less significant ones than other accounts do.

Moreover, I would suggest that the foregoing also diminishes the motivation for holding alternative views. For example, if we have an account that plausibly shows how moral facts reduce to recognisably natural facts, whilst explaining how we get an ‘ought’ from an ‘is’, and resisting the JTD objection and variants of the OQA, then the motivation for holding that moral facts and properties are autonomous from natural ones (as the non-naturalist proposes) is somewhat reduced. Further, if, in addition to avoiding a commitment to irreducibly normative facts and properties, this account successfully captures a very close connection between sincere moral judgements and motivation, then the impetus towards holding an expressivist view is lessened. And if our concept of a moral fact on this account is not a concept of an objectively and categorically prescriptive requirement (in Mackie’s sense) — so that these facts are not rendered metaphysically ‘queer’, and unknowable without some special faculty of moral perception or intuition — then some of the motivation for subscribing to a moral error theory is undercut.

There is room for reasonable debate about which challenges are the most important for Goal Theory to resist, and which theoretical adequacy criteria there are and what weight one should accord to them. Nevertheless, I would submit that any theory that plausibly satisfies all of the adequacy criteria that I surveyed, in addition to being resistant to the challenges that I did evaluate, would surely be a credible one, worthy of serious consideration by other philosophers. This is the category into which I now place Goal Theory, tentatively concluding that it is a plausible candidate for the novel but adequate theory that I described earlier.

I found Goal Theory to be adequate as a first-order moral theory too: explaining what is right and wrong (giving us a clear way of getting answers to our questions about actual moral situations); being comprehensive (giving us answers, or at least a way of establishing such answers, that we can imagine applying to any situation); being consistent (not yielding conflicting results in different circumstances); defusing possible conflicts between self-interest and morality; and explaining why we should be moral.

If we compare across the chapters, pulling together their themes and connecting their key messages, then what do we find? Well, in addition to accruing the general benefits of being a realist, reductive naturalist, representationalist, and cognitivist account, I would suggest the particular strength of my account lies in its distinctive reductionist characterisation, on which moral facts and properties are reductively identified with natural facts and properties of enlightened desire satisfaction (with the reductive identities in question ultimately being a matter of synthetic fact). This characterisation I find compelling because: (1) it was derived from my positive argument in section 2.1, rather than being created out of whole cloth, so if the argument is plausible then we have good reason to accept the characterisation; and (2) once conjoined with my related characterisations of normative reasons, inescapability, self-interest, and Moral Absolutism, it does much heavy lifting in terms of rendering my account theoretically adequate and resistant to dominant objections. Of particular note in this regard, we have the following:

- Since moral facts on Goal Theory are ultimately facts about agents' enlightened desires, then a distinctive internal connection between moral judgement and motivation is plausibly guaranteed by what the judgement is about (section 2.4).
- As moral facts and normative reasons on my account (with the latter defined by the HTR*) both derive from the serving of agents' enlightened desires, then

we find that agents will always have (decisive) reasons to comply with moral requirements (section 2.6).

- Combining the foregoing with my enlightened desire conception of Moral Absolutism, i.e. Moral Absolutism*, we find that there is no tension between the HTR*, Moral Rationalism, and Moral Absolutism* — thereby providing a plausible resolution to the Central Problem (section 3.1). Moreover, that moral facts and normative reasons on my account derive from the serving of agents' enlightened desires helps it to resist the overgeneration and undergeneration arguments (once I incorporate my proposed true strongest desire and my weighting scheme for reasons).
- On the conjunction of the HTR* and Goal Theory's particular enlightened desire conceptions, together with a related conception of inescapability, i.e. inescapability*, we find that categorical reasons may be denied, whilst still preserving a credible kind of 'practical clout', and avoiding disabling crucial uses of moral concepts (section 3.2).
- Again, on the conjunction of the HTR* and Goal Theory's particular enlightened desire conceptions, dedicated immoralists may have genuine reasons to refrain from their evil deeds, despite refraining serving none of their unenlightened desires — contra one interpretation of Shafer-Landau's argument (section 3.3). (The other interpretation is undone once we grant the unreliability of moral intuitions, per section 3.4.)
- The same conjunction (plus inescapability*) helps to ensure that the normative importance of normative facts and properties does not go missing in attempts to naturalise them, with morality then generating strong reasons to conform to

its standards, and normative reasons having the required normative importance or authority (section 4.5).

- My account already accrues general benefits from being an egoist one, viz. avoiding any possible conflict between self-interest and morality (since to act morally is always to act in one's self-interest), giving people a ready answer to the question of why they should be moral (because, on egoism, morality always best serves one's self-interest), and making moral behaviour rational by definition (on the assumption that it is rational to pursue one's own interests). However, once egoism is defined in terms of agents' strongest enlightened desires (i.e. egoism**), and combined with my proposed universal true strongest desire, then my account plausibly resists the standard internal and external objections to egoism (chapter 5).

Another significant finding to have emerged from my research is that while my account gives up on certain strongly-held notions (e.g. categorical reasons, the thought that the normative is just too different from the natural for the former to be a subset of the latter, and that we have natural moral duties to other people simply because our actions could help or harm them), my account still captures enough of the appearances that are insisted upon to be, overall, plausible.

What limitations must I acknowledge here? Well, most obviously, it was beyond the scope of my study to address every extant objection to accounts of Goal Theory's type. I adduced numerous positive arguments for Goal Theory's specific commitments, and answered a number of objections to its positive commitments and the arguments for them. However, while I tried to choose the most dominant and cogent objections, my selection clearly cannot be exhaustive. Moreover, for those objections to which I did respond, I could not look at every possible counterargument to my arguments. In this thesis, as in philosophy generally, there are always more arguments to be made.

As for other limitations, several cluster around the notion of agents' true strongest desires. In particular, I rely to varying degrees upon this desire being accessible (at least to some suitable level of approximation), to its being (almost) universal, to its being something in the region of a deep and abiding kind of satisfaction, and for the kinds of real-world actions that best serve this desire to approximately align with our stock moral truisms. I presented arguments for all of these dependencies, making the claims at least plausible, but a thorough (empirical) justification was beyond the scope of this study. Goal Theory would survive a failure in any of them, but its theoretical adequacy would be somewhat diminished.

Some opportunities for further research arise naturally from these limitations. In particular, the following are pertinent:

- I think a priority would be a cross-disciplinary research program, with philosophers collaborating with psychologists, neuroscientists, and others in an attempt to adequately specify and consequently discover individual, and potentially universal, true strongest desires in humans.
- This research on true strongest desires could then be usefully combined with research on the likely real-world consequences of certain (kinds of) actions, in an attempt to establish (universal) moral facts on Goal Theory.
- There would then be value in carrying out research into the possibility of genuine real-world exceptions to these moral facts, i.e. people whose true strongest desire would be best served by acting in conventionally immoral

ways (as with the hypothetical sensible knaves/dedicated immoralists that I spoke about in chapter 3) — either because they have an atypical true strongest desire, or because their individual circumstances are such that even the putative true strongest desire I posit would be best served for them by acting in this way.

- Based upon the foregoing, one might seek to establish if there are general ‘rules’ that should be instituted within Goal Theory.
- Moving away from metaethics and normative ethics, I think there would be significant value in establishing Goal Theory’s applied ethical positions, thereby enabling it to be utilised to help solve real-world ethical problems in areas such as bioethics, animal ethics, environmental ethics, and artificial intelligence. This might make Goal Theory a credible addition in this domain to the familiar triad of Utilitarianism, Kantianism, and virtue theory.
- Of course, there is scope for addressing objections to Goal Theory that had to be set aside here and engaging with any critical responses to Goal Theory that come from other philosophers.
- Finally, there would also be value in investigating how my account would handle ‘thick’ evaluative terms and concepts (which I set aside in my thesis), establishing if these have any bearing upon my answers to such questions as whether there is a fact-value distinction, whether there are ethical truths, and, if there are such truths, whether these truths are objective.

How much has the thesis moved the professional discussion along? As stated, Goal Theory was previously unknown in the academic literature, appearing only in two book chapters aimed at the educated lay reader, where it was of necessity presented in a relatively high-level fashion and left untested against common objections and unevaluated in terms of theoretical adequacy. In this thesis, I have presented Goal Theory in the more formal and detailed way that is appropriate to academic philosophy — unpacking the theory's metaphysical, epistemological, psychological, and other commitments, testing it against the dominant objections that are aimed at accounts of its type, and properly evaluating its theoretical adequacy.

In so doing, I have not merely defended Goal Theory itself, but have also plausibly solved several independent problems in metaethics and normative ethics. For example, whether a Humean account can survive the Central Problem, and plausibly retain crucial uses of moral concepts and yield reasons for dedicated immoralists to refrain from their evil deeds, whilst doing without categorical normative reasons; whether a naturalist account can show how such things as goodness and badness, rightness and wrongness, and reasons for acting, can all be captured entirely within a metaphysically naturalistic worldview; whether an egoist account can be internally coherent and not generate propositions that are false or unjustifiably unacceptable; and whether a non-analytic form of naturalism can survive the Moral Twin Earth argument. Along the way, I have made original developments to Goal Theory, including a 'rule' modification, and the creation of novel variants of the Humean Theory of Reasons and ethical egoism (with these improving upon existing variants).

Accordingly, in plausibly establishing that there is a novel theory that is both theoretically adequate and resistant to dominant challenges (and that also solves several independent problems), I think that this thesis has formed an original and substantial contribution to knowledge, constituting important progress in our quest to determine answers to the central metaethical and normative questions about the true nature of moral reality and the true content of morality. The potential implications of this research are substantial, in potentially altering the normative and metaethical landscapes, and yielding a number of counter-intuitive yet well-supported consequences.

While I acknowledge that I have not produced any *knockdown* arguments for Goal Theory or its components, I concur with Kagan's observation that:

> Almost any normative theory is likely to have its counterintuitive aspects, and people can sincerely disagree as to which theory is, on balance, the most attractive. That is why there are few or no 'knockdown' arguments in ethics (or anywhere, for that matter). All you can do is point out the attractive features of your own favored theory, explain why you are prepared to live with its various unattractive features, and try to show that the alternatives are even worse.[1]

In this thesis, I think I have presented a plausible, coherent, and compelling case with which other philosophers might engage; and their doing so would help to raise the level of argumentation and analysis, enabling Goal Theory in particular, and metaethics and normative ethics in general, to be usefully refined and developed. Even if Goal Theory ultimately succumbs to some other objection, I think that this process of engagement and development will still be a useful one for philosophy. Therefore, to other philosophers I offer a conciliatory invitation to be open-minded about Goal Theory, to engage with it as seriously and charitably as they do with other comparable theories, and to contribute to the discussion that I have initiated here. Not least, I hope that I have made good on my intention that the outcome of my research should be both highly surprising and resistant to easy refutation.

[1] Kagan, *Normative Ethics*, p. 16.

Bibliography

Alexander, Joshua , and Weinberg, Jonathan M. , 'The “Unreliability” of Epistemic Intuitions', in *Current Controversies in Experimental Philosophy*, ed. by E. Machery and E. O’Neill (Oxford: Routledge, 2014), pp. 128-45.

Alexander, Joshua, Mallon, Ronald, and Weinberg, Jonathan M. , 'Accentuate the Negative', *Review of Philosophy and Psychology,* 2 (2010), 297–314.

Alexander, Joshua, and Weinberg, Jonathan M., 'Analytic Epistemology and Experimental Philosophy', *Philosophy Compass*, 2 (2007), 56–80.

Audi, Robert, 'Intuition, Inference, and Rational Disagreement in Ethics', *Ethical Theory and Moral Practice*, 11 (2008), 475–92.

Axelrod, Robert, 'The Emergence of Cooperation among Egoists', *The American Political Science Review* (1981), 306-18.

———, *The Evolution of Cooperation* (New York: Basic Books, 1984).

Ayer, A.J., *Language, Truth and Logic*, 2nd edn (London: Gollancz, 1946[1936]).

Baier, Kurt, *The Moral Point of View* (Ithaca, NY: Cornell University Press, 1958).

Barker, S., 'Is Value Content a Component of Conventional Implicature?', *Analysis* (2000), 268–79.

Baumeister, R. F. , and Leary, M. R. , 'The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation', *Psychological Bulletin*, 117 (1995), 497–529.

Bealer, G., 'Intuition and the Autonomy of Philosophy', in *Rethinking Intuition*, ed. by M. DePaul and W. Ramsey (Lanham, Md.: Rowman & Littlefield, 1998), pp. 201-40.

Besser-Jones, Lorraine 'Personal Integrity, Morality and Psychological Well-Being: Justifying the Demands of Morality', *Journal of Moral Philosophy* (2008), 361–83.

Bicchieri, Cristina 'Rationality and Game Theory', in *The Handbook of Rationality* (Oxford: Oxford University Press, 2003).

Binmore, Ken, *Natural Justice* (Oxford: Oxford University Press, 2005).

Binmore, Kenneth, *Playing Fair: Game Theory and the Social Contract*, Vol. 1 (Cambridge, MA: MIT Press, 1994).

Blackburn, Simon, *Spreading the Word* (Oxford: Oxford University Press, 1984).

BonJour, L., *In Defense of Pure Reason* (Cambridge: Cambridge University Press, 1998).

Boyd, Richard, 'How to Be a Moral Realist', in *Essays on Moral Realism*, ed. by G. Sayre-McCord (Ithaca, NY: Cornell University Press, 1988), pp. 181-228.

Brandt, Richard, *A Theory of the Good and the Right* (New York: Oxford University Press, 1979).

———, 'The Science of Man and Wide Reflective Equilibrium', *Ethics*, 100 (1990), 259–78.

Brink, David, *Moral Realism and the Foundation of Ethics* (Cambridge: Cambridge University Press, 1989).

———, *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press, 1989).

Buchanan, Allen, *Ethics, Efficiency, and the Market* (Totowa, NJ: Rowman & Allanheld, 1985).

Buckwalter, Wesley, and Stich, Stephen, 'Gender and Philosophical Intuition', in *Experimental Philosophy*, ed. by Joshua Knobe and Shaun Nichols (Oxford: Oxford University Press, 2014).

Bukoski, M., 'A Critique of Smith's Constitutivism', *Ethics* (2016), 116–46.

Burgess-Jackson, Keith, 'Taking Egoism Seriously', *Ethical Theory and Moral Practice,* 16 (2013), 529-42.

Carbonell, Vanessa, 'De Dicto Desires and Morality as Fetish', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* 163 (2013), 459-77.

Carrier, Richard, *Sense and Goodness without God: A Defense of Metaphysical Naturalism* (Bloomington, IN: AuthorHouse, 2005).

———, 'On Defining Naturalism as a Worldview', *Free Inquiry,* 30 (2010), 50-51.

———, 'Moral Facts Naturally Exist (and Science Could Find Them)', in *The End of Christianity,* ed. by John Loftus (Amherst, NY: Prometheus, 2011), pp. 333-64.

———, *Proving History: Bayes’s Theorem and the Quest for the Historical Jesus* (Amherst, NY: Prometheus Books, 2012).

Copp, D., 'Milk, Honey, and the Good Life on Moral Twin Earth', *Synthese,* 124 (2000), 113-37.

———, 'Normativity and Reasons: Five Arguments from Parfit against Normative Naturalism', in *Ethical Naturalism: Current Debates*, ed. by Susana Nuccetelli and Gary Seay (Cambridge: Cambridge University Press, 2011), pp. 24-57.

Cornman, J.W., Lehrer, K., and Pappas, G.S., *Philosophical Problems and Arguments: An Introduction*, 3rd edn (New York: Macmillan Publishing Company, 1982).

Cosmides, L., and Tooby, J., 'Cognitive Adaptations for Social Exchange', in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, ed. by J.H. Barkow, L. Cosmides and J. Tooby (Oxford: Oxford University Press, 1992).

Crisp, R., 'Naturalism and Non-Naturalism in Ethics', in *Identity, Truth and Value*, ed. by S. Lovibond and S.G. Williams (Malden, MA: Blackwell Publishers, 1996), pp. 113–29.

Cummins, Robert, 'Reflections on Reflective Equilibrium', in *Rethinking Intuition*, ed. by M. DePaul and W. Ramsey (Lanham, Md: Rowman & Littlefield, 1998), pp. 113-28.

Cuneo, Terence, 'Moral Naturalism and Categorical Reasons', in *Ethical Naturalism: Current Debates*, ed. by Susana Nuccetelli and Gary Seay (Cambridge: Cambridge University Press, 2011), pp. 110-30.

D'Agostini, Giulio, 'Role and Meaning of Subjective Probability: Some Comments on Common Misconceptions', in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by Ali Mohammad-Djafari (Melville, NY: American Institute of Physics, 2001), pp. 23-30.

Dancy, J., *Practical Reality* (Oxford: Oxford University Press, 2000).

———, 'Nonnaturalism', in *The Oxford Handbook of Ethical Theory*, ed. by D. Copp (Oxford: Oxford University Press, 2005).

Daniels, N., 'Wide Reflective Equilibrium and Theory Acceptance in Ethics', *The Journal of Philosophy*, 76 (1979), 256– 82.

Daniels, Norman, 'Wide Reflective Equilibrium and Theory Acceptance in Ethics', *Journal of Philosophy*, 76 (1979), 256-82.

Darwall, Stephen, *Impartial Reason* (Ithaca, NY: Cornell University Press, 1983).

De Brigard, F., 'If You Like It, Does It Matter If It's Real?', *Philosophical Psychology*, 23 (2010), 43–57.

De Waal, Frans, *Primates and Philosophers: How Morality Evolved*, ed. by Stephen Macedo and Josiah Ober (Princeton, NJ: Princeton University Press, 2009).

Donagan, A., 'W. A. Frankena and G. E. Moore's Metaethics', *Monist* (1981), 293-304.

Doris, John M. , and Plakias, Alexandra, 'How to Argue About Disagreement: Evaluative Diversity and Moral Realism', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*, ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2007).

Dowell, J., 'The Metaethical Insignificance of Moral Twin Earth', in *Oxford Studies in Metaethics (Vol. 11)*, ed. by R. Shafer-Landau (Oxford: Oxford University Press, 2015).

Drescher, Gary L., *Good and Real: Demystifying Paradoxes from Physics to Ethics* (Cambridge, MA: The MIT Press, 2006).

Enoch, David, 'How Is Moral Disagreement a Problem for Realism?', *The Journal of Ethics*, 13 (2009), 15-50.

———, *Taking Morality Seriously* (Oxford: Oxford University Press, 2011).

Feldman, Fred, *Introductory Ethics* (London: Pearson, 1978).

Finlay, S., 'The Reasons That Matter', *Australasian Journal of Philosophy* (2006), 1-20.

———, *Confusion of Tongues: A Theory of Normative Language* (Oxford: Oxford University Press, 2014).

FitzPatrick, W., 'Skepticism About Naturalizing Normativity: In Defense of Ethical Nonnaturalism', *Res Philosophica* (2014), 559–88.

Fitzpatrick, W. J., 'Robust Ethical Realism, Non-Naturalism and Normativity', *Oxford Studies in Metaethics* (2008), 159-206.

———, 'Ethical Non-Naturalism and Normative Properties', in *New Waves in Metaethics*, ed. by M. Brady (Basingstoke: Palgrave Macmillan, 2010), pp. 7-35.

Foot, Philippa, 'Morality as a System of Hypothetical Imperatives', *Philosophical Review*, 81 (1972), 305-15.

———, *Natural Goodness* (Oxford: Oxford University Press, 2001).

Gampel, Eric H., 'A Defense of the Autonomy of Ethics: Why Value Is Not Like Water', *Canadian Journal of Philosophy*, 26 (1996), 191-209.

Gampel, Eric H. , 'Ethics, Reference, and Natural Kinds', *Philosophical Papers* (1997), 147-63.

Geach, P., 'Assertion', *Philosophical Review* (1964), 449-65.

Gert, Bernard, *Morality: Its Nature and Justification, Revised Edition* (Oxford: Oxford University Press, 2005).

———, 'Morality', in *The Cambridge Dictionary of Philosophy*, ed. by Robert Audi (Cambridge: Cambridge University Press, 2015).

Gibbard, Alan, *Wise Choices, Apt Feelings* (Oxford: Clarendon Press, 1990).

Gowans, Christopher, 'Introduction', in *Moral Disagreements: Classic and Contemporary Readings*, ed. by Christopher Gowans (2000), pp. 1-43.

Greene, J., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L., and Cohen, J., 'Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment', *Cognition*, 111 (2009), 364– 71.

Grossmann, Reinhardt, *The Categorial Structure of the World* (Bloomington, Indiana: Indiana University Press, 1983).

Haidt, J., and Baron, J., 'Social Roles and the Moral Judgement of Acts and Omissions', *European Journal of Social Psychology*, 26 (1996), 201– 18.

Haidt, J., and Joseph, C, 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues', *Daedalus: On Human Nature*, 133 (2004), 55-66.

Haidt, Jonathan, 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', *Psychological Review*, 108 (2001), 814-34.

Haidt, Jonathan, and Joseph, Craig, 'Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues', *Daedalus*, 133 (2004), 55–66.

Hardin, Garrett, 'The Tragedy of the Commons', *Science*, 162 (1968), 1243-48.

Hare, R.M., 'Rawls' Theory of Justice', *Philosophical Quarterly*, 23 (1973), 144–55, 241–51.

Harman, Gilbert, 'Moral Relativism Defended', *Philosophical Review*, 85 (1975), 3-22.

———, 'Ethics and Observation', in *The Nature of Morality: An Introduction to Ethics* (Oxford: Oxford University Press, 1977), pp. 3-10.

Hart, H.L.A., *The Concept of Law*, 3rd edn (Oxford: Oxford University Press, 2012 [1961]).

Hauser, M. D., *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong* (New York: Ecco Press, 2006).

Hechter, Michael, and Opp, Karl-Dieter, *Social Norms* (New York: Russell Sage Foundation, 2005).

Hempel, C., 'Comments on Goodman's Ways of Worldmaking', *Synthese*, 45 (1980), 193-99.

Hoffman, Joshua, and Rosenkrantz, Gary S., *Substance among Other Categories* (Cambridge: Cambridge University Press, 1994).

Horgan, T., and Graham, G., 'In Defense of Southern Fundamentalism', *Philosophical Studies* (1991), 107-34.

Horgan, T., and Timmons, M., 'Troubles for New Wave Semantics', *Philosophical papers* (1992), 153-75.

———, 'Copping out on Moral Twin Earth', *Synthese*, 124 (2000), 139-52.

Horgan, Terence, and Timmons, Mark, 'Analytic Moral Functionalism Meets Moral Twin Earth', in *Minds, Ethics, and Conditionals: Themes from the Philosophy of Frank Jackson*, ed. by Ian Ravenscroft (Oxford: Oxford University Press, 2009), pp. 221-36.

Horgan, Terry, and Timmons, Mark, 'New Wave Moral Realism Meets Moral Twin Earth', *Philosophy and Phenomenological Research*, 16 (1991), 447-65.

Horwich, P., 'Gibbard's Theory of Norms', *Philosophy and Public Affairs* (1993), 67–79.

Hospers, John, 'Baier and Medlin on Ethical Egoism', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 12 (1961), 9-16.

Hume, David, *A Treatise of Human Nature* (Oxford: Clarendon Press, 1888 [1968]).

———, 'An Enquiry Concerning the Principles of Morals', in *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. by L.A. Selby-Bigge and P.H. Nidditch (Oxford: Clarendon, 1975 [1777]).

Jackson, F., 'A Problem for Expressivism', *Analysis* (1998), 239–51.

Jackson, Frank, *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press, 1998).

Johnston, M., 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society (Supp.)* (1989), 139-74.

Joyce, James, 'Bayes' Theorem', *The Stanford Encyclopedia of Philosophy*, Winter 2016 edn, ed. by Edward N. Zalta <https://plato.stanford.edu/archives/win2016/entries/bayes-theorem/> [accessed 6th March 2019].

Joyce, Richard, *The Myth of Morality* (Cambridge: Cambridge University Press, 2001).

———, *The Evolution of Morality* (Cambridge, MA: MIT Press, 2006).

———, 'Metaethical Pluralism: How Both Moral Naturalism and Moral Skepticism May Be Permissible Positions', in *Ethical Naturalism: Current Debates*, ed. by Susana Nuccetelli and Gary Seay (Cambridge: Cambridge University Press, 2012), pp. 89-109.

———, 'Error Theory', in *International Encyclopedia of Ethics*, ed. by H. LaFollette (Hoboken, NJ: Wiley-Blackwell, 2013).

Kagan, Shelley, *Normative Ethics* (Boulder, CO: Westview Press, 1998).

———, 'Thinking About Cases', in *Moral Knowledge*, ed. by Ellen Frankel Paul, Jr. Miller, Fred and Jeffrey Paul (Cambridge: Cambridge University Press, 2001), pp. 44-63.

Kahneman, Daniel, and Deaton, Angus 'High Income Improves Evaluation of Life but Not Emotional Well-Being', *Proceedings of the National Academy of Sciences*, 7 (2010), 16489–93.

Kavka, G.S., *Hobbesian Moral and Political Theory* (Princeton, NJ: Princeton University Press, 1986).

Kifer, Yona , Heller, Daniel , Perunovic, Wei Qi Elaine, and Galinsky, Adam D. , 'The Good Life of the Powerful: The Experience of Power and Authenticity Enhances Subjective Well-Being', *Psychological Science*, 24 (2013), 280-88.

Kim, J., 'The Myth of Nonreductive Materialism', *Proceedings and Addresses of the American Philosophical Association*, 63 (1989), 31-47.

Kirchin, Simon (ed.), *Thick Concepts* (Oxford: Oxford University Press, 2013).

Kitcher, Philip, 'Biology and Ethics', in *The Oxford Handbook of Ethical Theory*, ed. by D. Copp (Oxford: Oxford University Press, 2005).

———, 'Between Fragile Altruism and Morality: Evolution and the Emergence of Normative Guidance', in *Evolutionary Ethics and Contemporary Biology*, ed. by G. Boniolo and G. De Anna (Cambridge: Cambridge University Press, 2006), pp. 159–77.

Kornblith, H., 'Appeals to Intuition and the Ambitions of Epistemology', in *Epistemology Futures*, ed. by S. Hetherington (Oxford: Oxford University Press, 2006), pp. 10-25.

Korsgaard, Christine, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).

Kripke, Saul, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980).

Lewis, D., 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61 (1983), 343–77.

Lewis, David, 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society (Supp.)*, 63 (1989), 113-37.

Lipton, P., *Inference to the Best Explanation* (London: Routledge, 2004).

Little, M., 'Moral Realism 2: Non-Naturalism', *Philosophical Books*, 35 (1994), 225–32.

Loeb, D., 'Moral Realism and the Argument from Disagreement', *Philosophical Studies* (1998), 281-303.

Lowe, E.J., *The Four-Category Ontology: A Metaphysical Foundation for Natural Science* (Oxford: Clarendon Press, 2006).

Luco, Andrés Carlos, 'Non-Negotiable: Why Moral Naturalism Cannot Do Away with Categorical Reasons', *Philosophical Studies*, 173 (2016), 2511–28.

Mackie, J.L., *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin, 1977).

McDowell, John, *Mind, Value, and Reality* (Cambridge, MA: Harvard University Press, 1998).

McGinn, C., *Ethics, Evil, and Fiction* (Oxford: Clarendon Press, 1997).

McNaughton, D., and Rawling, P., 'Naturalism and Normativity', *Supplement to the Proceedings of the Aristotelian Society* (2003), 23–45.

McPherson, Tristram, 'Ethical Non-Naturalism and the Metaphysics of Supervenience', in *Oxford Studies in Metaethics*, ed. by R. Shafer-Landau (Oxford: Oxford University Press, 2012), pp. 205-34.

———, 'Review: Mark Schroeder's Hypotheticalism: Agent-Neutrality, Moral Epistemology, and Methodology', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 157 (2012), 445-53.

Miller, Alexander, *Contemporary Metaethics: An Introduction*, 2nd edn (Cambridge: Polity Press, 2013).

Millgram, E., 'Was Hume a Humean?', *Hume Studies*, 31 (1995), 75-93.

Moore, G.E., *Principia Ethica*, revised edn (Cambridge: Cambridge University Press, 1993[1903]).

Nagel, Thomas, *The View from Nowhere* (Oxford: Oxford University Press, 1986).

Nash, John, 'Equilibrium Points in N-Person Games', *Proceedings of the National Academy of Sciences*, 36 (1950), 48-49.

Nisbett, R. E., *The Geography of Thought: How Asians and Westerners Think Differently...And Why* (New York: Free Press, 2003).

Nozick, Robert, *Anarchy, State, and Utopia* (New York: Basic Books, 1974).

———, *The Examined Life* (New York: Simon & Schuster, 1989).

Okasha, S., 'Van Fraassen’s Critique of Inference to the Best Explanation', *Studies in History and Philosophy of Science* (2000), 691–710.

Olson, Jonas, 'Are Desires De Dicto Fetishistic?', *Inquiry*, 45 (2002), 89-96.

Ophuls, W., 'Leviathan or Oblivion', in *Toward a Steady State Economy* (San Francisco, CA: Freeman, 1973), pp. 215-30.

Ostrom, Elinor, *Governing the Commons: The Evolution of Institutions for Collective Action (Political Economy of Institutions and Decisions)* (Cambridge: Cambridge University Press, 1990).

Owen, Jonathan, and Gray, Sadie, 'Ronnie Biggs Pleads: Let Me out So I Can Die with My Family', *Independent on Sunday*, 30th December 2007 30th December 2007.

Paakkunainen, Hille, 'The "Just Too Different" Objection to Normative Naturalism', *Philosophy Compass* (forthcoming).

Parfit, Derek, *On What Matters Vol 1* (Oxford: Oxford University Press, 2011).

———, *On What Matters Vol 2* (Oxford: Oxford University Press, 2011).

Piliavin, J. A., 'Doing Well by Doing Good: Benefits for the Benefactor', in *Flourishing: Positive Psychology and the Life Well-Lived*, ed. by C. L. M. Keyes and J. Haidt (Washington, DC: American Psychological Association, 2003), pp. 227– 47.

Pinker, Steven, *The Better Angels of Our Nature: The Decline of Violence in History and Its Causes* (London: Penguin Books, 2011).

Pizarro, David A., Tannenbaum, David, Uhlmann, Eric Luis, and Ditto, Peter H., 'The Motivated Use of Moral Principles', *Judgment and Decision Making*, 4 (2009), 476-91.

Post, Stephen G., 'Altruism, Happiness, and Health: It's Good to Be Good', *International Journal of Behavioral Medicine*, 12 (2005), 66–77.

Price, Richard, 'A Review of the Principle Questions in Morals', in *The British Moralists 1650–1800, Ii*, ed. by D.D. Raphael (Oxford: Clarendon Press, 1758/1969), pp. 131–98.

Prinz, Jesse, *The Emotional Construction of Morals* (Oxford: Clarendon Press, 2007).

Putnam, Hilary, 'The Meaning of 'Meaning'', in *Language, Mind and Knowledge*, ed. by Keith Gunderson (Minneapolis, MN: University of Minnesota Press, 1975).

Rachels, James, 'Two Arguments against Ethical Egoism', *Philosophia* (1974), 297-314.

———, 'Ethical Egoism', in *Ethical Theory: An Anthology*, ed. by Russ Shafer-Landau (Hoboken, NJ: John Wiley and Sons, 2012), pp. 193-99.

Railton, Peter, 'Moral Realism', *Philosophical Review*, 95 (1986), 163-207.

———, 'Naturalism and Prescriptivity', *Social Philosophy and Policy*, 7 (1989), 151-74.

———, 'Humean Theory of Practical Rationality', in *The Oxford Handbook of Ethical Theory*, ed. by David Copp (Oxford: Oxford University Press, 2006), pp. 265-81.

Rawls, J., 'Outline of a Decision Procedure for Ethics', *The Philosophical Review*, 60 (1951), 177– 97.

Raz, J., *Engaging Reason: On the Theory of Value and Action* (Oxford: Oxford University Press, 1999).

Ridge, M., 'Ecumenical Expressivism: Finessing Frege', *Ethics*, 116 (2006), 302–36.

Ridley, M., *The Origins of Virtue: Human Instincts and the Evolution of Cooperation* (London: Penguin, 1997).

Rolston, H., *Environmental Ethics: Duties to and Values in the Natural World* (Philadelphia, PA: Temple University Press, 1988).

Roskies, A., 'Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy."', *Philosophical Psychology*, 16 (2003), 51-66.

Ross, W.D., *The Right and the Good* (Oxford: Clarendon Press, 1930/2002).

Scanlon, T. M., *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998).

Schroeder, M., 'Realism and Reduction: The Quest for Robustness', *Philosophers' Imprint*, 5 (2005), 1-18.

———, 'Hybrid Expressivism: Virtues and Vices', *Ethics*, 119 (2009), 257–309.

Schroeder, Mark, *Slaves of the Passions* (Oxford: Oxford University Press, 2007).

Shafer-Landau, R., *Moral Realism: A Defense* (Oxford: Oxford University Press, 2003).

———, 'Defending Ethical Intuitionism', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*, ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2008), pp. 83-96.

———, 'A Defence of Categorical Reasons', *Proceedings of the Aristotelian Society (Supp.)*, 109 (2009), 189-206.

———, 'Review: Three Problems for Schroeder's Hypotheticalism', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 157 (2012), 435-43.

Shafer-Landau, Russ, 'Error Theory and the Possibility of Normative Ethics', *Philosophical Issues* (2005), 107-20.

Shafer-Landau, Russ, and Cuneo, Terence, *Foundations of Ethics: An Anthology*, ed. by Russ Shafer-Landau and Terence Cuneo (Oxford: Blackwell Publishing Ltd, 2007).

Shtulman, Andrew, and Harrington, Kelsey, 'Tensions between Science and Intuition across the Lifespan', *Topics in Cognitive Science* (2016), 118–37.

Shtulman, Andrew, and Valcarcel, Joshua, 'Scientific Knowledge Suppresses but Does Not Supplant Earlier Intuitions', *Cognition* (2012), 209-15.

Sidgwick, H., *The Methods of Ethics*, 7th edn (Indianapolis, IL: Hackett Publishing Company, 1981 (1907)).

Singer, Peter, 'Sidgwick and Reflective Equilibrium', *Monist* (1974), 490-517.

Sinhababu, Neil, 'The Humean Theory of Motivation Reformulated and Defended', *Philosophical Review*, 118 (2009), 465–500.

———, 'Ethical Reductionism', <http://philpapers.org/archive/SINER.pdf>.

Sinnott-Armstrong, Walter, 'Moral Intuitionism Meets Empirical Psychology', in *Metaethics after Moore*, ed. by Terry Horgan and Mark Timmons (Oxford: Oxford University Press, 2006), pp. 339-66.

———, 'Framing Moral Intuitions', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*, ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2007).

———, 'How to Apply Generalities: Reply to Tolhurst and Shafer-Landau', in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*, ed. by Walter Sinnott-Armstrong (Cambridge, MA: Bradford Books, 2007).

———, *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, 4 vols, Vol. 3 (Cambridge, MA: MIT Press, 2007).

———, *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*, 4 vols, Vol. 1 (Cambridge, MA: MIT Press, 2008), pp. 1-46.

Sinnott-Armstrong, Walter, Young, Liane, and Cushman, Fiery, 'Moral Intuitions', in *The Moral Psychology Handbook*, ed. by John M. Doris (Oxford: Oxford University Press, 2010), pp. 246-72.

Smart, JJC, 'An Outline of a System of Utilitarian Ethics', in *Utilitarianism: For and Against*, ed. by JJC Smart and B. Williams (Cambridge: Cambridge University Press, 1973), pp. 1-74.

Smith, Michael, *The Moral Problem* (Oxford: Blackwell, 1994).

———, 'In Defense of "the Moral Problem": A Reply to Brink, Copp, and Sayre-Mccord', *Ethics*, 108 (1997), 84-119.

———, 'Moral Realism', in *The Blackwell Guide to Ethical Theory*, ed. by H. LaFollette (Oxford: Blackwell, 2000), pp. 15-37.

———, 'A Constitutivist Theory of Reasons: Its Promise and Parts', *Law, Ethics and Philosophy* (2013), 9-30.

———, 'The Magic of Constitutivism', *American Philosophical Quarterly* (2015), 187-200.

Sober, E., and Wilson, D. S., *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998).

Sommers, F., 'Types and Ontology', *Philosophical Review* (1963).

Sosa, Ernest, 'Experimental Philosophy and Philosophical Intuition', *Philosophical Studies*, 132 (2007), 99-107.

Stevenson, Charles L. , 'The Nature of Ethical Disagreement', in *Exploring Philosophy: An Introductory Anthology*, ed. by Steven M. Cahn (Oxford: Oxford University Press, 2009 [1963]).

Stoljar, D., 'Emotivism and Truth Conditions', *Philosophical Studies* (1993), 81–101.

Stratton-Lake, P., 'Self-Evidence, Intuition and Understanding', in *Madison Workshop in Metaethics* (Madison, WI: 2016).

Strawson, P.F., 'Categories', in *Ryle: A Collection of Critical Essays*, ed. by O.P. Wood and G. Pitcher (London: Macmillan, 1970).

Street, S., 'A Darwinian Dilemma for Realist Theories of Value', *Philosophical Studies*, 127 (2006), 109–66.

Sturgeon, Nicholas, 'Moral Explanations', in *Essays on Moral Realism*, ed. by G. Sayre-McCord (Ithaca, NY: Cornell University Press, 1988), pp. 229-56.

———, 'Ethical Naturalism', in *Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press, 2006), pp. 91–121.

Sumner, L. W., 'Welfare, Happiness, and Pleasure', *Utilitas*, 4 (1992), 199–223.

Swain, Stacey, Alexander, Joshua, and Weinberg, Jonathan M., 'The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp', *Philosophy and Phenomenological Research*, 76 (2008), 138-55.

Talbot, Brian, 'Psychology and the Use of Intuitions in Philosophy', *Studia Philosophica Estonica*, 2 (2009), 157-76.

Tersman, Folke, *Moral Disagreement* (Cambridge: Cambridge University Press, 2006).

Thoits, P. A., and Hewitt, L. N. , 'Volunteer Work and Well-Being', *Journal of Health and Social Behavior*, 42 (2001), 115– 31.

Timmons, Mark, *Morality without Foundations: A Defense of Ethical Contextualism* (Oxford: Oxford University Press, 1999).

Tolhurst, W., 'The Argument from Moral Disagreement', *Ethics* (1987), 610-21.

Tresan, Jon, 'Question Authority: In Defense of Moral Naturalism without Clout', *Philosophical Studies* (2010), 221-38.

Trivers, Robert L., 'The Evolution of Reciprocal Altruism', *The Quarterly Review of Biology,* 46 (1971), 35-57.

Tucker (ed.), Chris, *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism* (Oxford: Oxford University Press, 2013).

Tversky, A., and Kahneman, D., 'The Framing of Decisions and the Psychology of Choice', *Science,* 211 (1981), 453– 58.

Van Deemter, Kees, *Not Exactly: In Praise of Vagueness* (Oxford: Oxford University Press, 2010).

van Roojen, M., 'Expressivism and Irrationality', *Philosophical Review* (1996), 311-55.

———, 'Expressivism, Supervenience and Logic', *Ratio* (2005), 190-205.

Weijers, D., 'Nozick's Experience Machine Is Dead, Long Live the Experience Machine!', *Philosophical Psychology*, 27 (2014), 513–35.

Weinberg, J., Nichols, S., and Stich, S., 'Normativity and Epistemic Intuitions', *Philosophical Topics* (2001), 429-60.

Weinstein, N., and Ryan, R. M. , 'When Helping Helps: Autonomous Motivation for Prosocial Behavior and Its Influence on Well-Being for the Helper and Recipient', *Journal of Personality and Social Psychology*, 98 (2010), 222– 44.

Weisberg, J., 'Locating Ibe in the Bayesian Framework', *Synthese* (2009), 125–44.

White, Matthew, *The Great Big Book of Horrible Things: The Definitive Chronicle of History's 100 Worst Atrocities* (New York: W. W. Norton & Company, 2011).

Wiggins, David, 'A Neglected Position?', in *Reality, Representation, and Projection*, ed. by John Haldane and Crispin Wright (Oxford: Oxford University Press, 1993), pp. 329-36.

Wong, David, 'On Moral Realism without Foundation', *The Southern Journal of Philosophy* (1986), 95-113.

———, *Natural Moralities: A Defense of Pluralistic Relativism* (Oxford: Oxford University Press, 2006).

Zangwill, N., 'Moral Modus Ponens', *Ratio* (1992), 177-93.