FIDUCIARY NON-COGNITIVISM

ADAM BUGEJA

A thesis submitted for the degree of
Doctor of Philosophy of the Australian National University

December 2017

©Copyright by Adam Bugeja 2017
This dissertation is solely the work of its author. No part of it has previously been submitted for any other degree, or is currently being submitted for any other degree. To the best of my knowledge, any help received in preparing this dissertation, and all sources used, have been duly acknowledged.

Adam Bugeja, 8 December 2017
Acknowledgements

I wish to thank my supervisory panel, Christian Barry, Seth Lazar, and Daniel Stoljar, and especially my primary supervisor Nicholas Southwood, for their advice and helpful criticisms. I am particularly grateful to Nic for his patience in letting me pursue this large project, even though it meant a change of topic in the fifth year of my candidature and numerous extension applications. Thanks to Robert Goodin, my primary supervisor in the first year of my candidature, one of whose reading recommendations provided the inspiration for Chapter 9 and my first publication.

Thanks to the audiences of a 2016 PhilSoc seminar in which I presented material from Chapters 1 and 5; a 2013 AAP presentation at the University of Queensland in which I presented a version of Chapter 9; and a 2012 MSPT Workshop on a paper in which I developed ideas that now appear in Chapter 8. Chapter 9 also benefited from the scrutiny and suggestions of an anonymous reviewer for Philosophical Studies and two anonymous reviewers for Ethics.

For the three years I spent working on ‘fiduciary non-cognitivism’ I made extensive use of the writing events run, and writing groups organized, by ANU Research Skills and Training, and funded by the Postgraduate and Research Students’ Association (PARSA). I am grateful for these services, and wish to thank the Research Training Community Manager Victoria Firth-Smith for her work and assistance.
Abstract

I present and defend a novel non-cognitivist theory of our moral thought and practices. It holds that moral judgements depend on the existence of broadly ‘social’ relationships among those who make them (among ‘moralists’, as I shall say).

I am led to this position by observations about non-cognitive disagreement. Theorists can satisfactorily explain the phenomenon of moral disagreement as a kind of non-cognitive disagreement only if they predict that it will have a property which I call ‘robustness’. Roughly, a candidate disagreement is robust if its status as a disagreement is difficult to dispute. I argue that parties to a robust non-cognitive disagreement necessarily share a type of trust relationship. One party entrusts to the other a concern to accommodate a certain subset of her desires in his decisions. The ‘entrusted party’ can resist the relevant desires of the ‘entrusting party’ without violating her trust only if satisfying them would frustrate a certain subset of his desires. I call this an entrusted concern relationship. Robust non-cognitive disagreement exists when the parties to such a relationship have opposing preferences, based on their relevant desires, concerning how the entrusted party is to act. I call such preferences ‘voiceable preferences’.

To explain moral disagreement and predict its robustness, non-cognitivists must posit the existence of entrusted concern relationships linking all moralists. I call the species of non-cognitivism which holds that the non-cognitive components of moral judgements are the voiceable preferences of parties to a network of entrusted concern relationships linking all moralists (a ‘moral trust network’) fiduciary non-cognitivism. This dissertation develops a non-cognitivist theory of this sort. I describe a form the moral trust network might take that would explain various elementary features of moral thought and talk, and defend the postulation of the described network.

I argue that this postulation is not too extravagant by explaining how and why such a network might exist, and by giving plausible examples of smaller trust networks with similar properties. I also contend, following Philippa Foot, that approval and disapproval presuppose the existence of a social context entitling approving parties to voices in approved parties’ decisions, and argue that, in light of this, we need to posit something like a moral trust network to provide a wholly satisfactory explanation of moral approval and disapproval.
Beyond moral disagreement and approval, my non-cognitivist theory offers at least partial explanations for the following phenomena: the use of moral judgements to justify and criticize actions; the rule-like character of moral requirements and the desirability of simplicity in moral theorizing; the changes of moral motivation that sometimes result from moral reasoning; moral uncertainty and its typical practical consequences; and the intuition that it can seem very odd to report that one has forgotten a moral truth. The theory’s explanations of features of moral thought and talk are fairly unified: they mostly derive from postulations about the shared ends of moralists, and the idea that the ‘terms’ of the relationships in the moral trust network would be shaped to serve these ends.
Contents

0 Introduction ........................................................................................................................................ 1
  0.1 Non-cognitivism and its Place in Metaethics ........................................................................ 2
  0.1.1 Distinguishing Non-cognitivism from Cognitivism ..................................................... 2
  0.1.2 Prominent Contemporary Non-cognitivist Theories ................................................. 6
  0.1.3 Standard Arguments For and Against Non-cognitivism ........................................... 7
  0.2 The Project ............................................................................................................................. 11
  0.2.1 Metaethical Desiderata that are Relevant to the Project ........................................... 11
  0.2.2 The Central Idea of Fiduciary Non-cognitivism ......................................................... 18
  0.2.3 Fiduciary Non-cognitivism and the Nine Desiderata ................................................. 21
  0.2.4 Scope and Omissions ...................................................................................................... 27
  0.2.5 Method ............................................................................................................................ 30
  0.2.6 A Look Ahead .................................................................................................................. 31

Part I Trust and Non-cognitive Disagreement ........................................................................... 33

1 Conditions for Robust Non-cognitive Disagreement ................................................................. 34
  1.1 The Phenomenon and Some Explanations ....................................................................... 35
  1.1.1 Robust Non-cognitive Disagreement ....................................................................... 35
  1.1.2 Type-of-Attitude Explanations ................................................................................ 37
  1.1.3 Attitudes, Plus Attitudes about Attitudes ................................................................. 39
  1.1.4 Is ‘Robust Non-cognitive Disagreement’ Really Non-Cognitive? ............................ 40
  1.1.5 Ridge’s ‘Disagreement in Prescription’ ................................................................... 43
  1.1.6 A Guiding Thought: The Reason-to-Consider Hypothesis ....................................... 45
  1.1.7 The Coordination Hypothesis ................................................................................... 48
  1.2 The Entrusted Concerns Hypothesis ............................................................................... 50
  1.2.1 Relevant Features of John’s Relationship with his Mother ................................... 51
  1.2.2 Entrusted Concern Relationships .......................................................................... 53
  1.2.3 A Statement of the Entrusted Concerns Hypothesis .............................................. 58
  1.2.4 Predictions of the Entrusted Concerns Hypothesis in Other Cases ..................... 60
  1.2.5 Are Cases of Imperative Disagreement Counterexamples to the Hypothesis? .... 63

2 Third-Party Non-cognitive Disagreement ............................................................................... 70
  2.1 Identifying Third-Party Disagreements ......................................................................... 71
  2.2 A Theory of Third-Party Disagreement ....................................................................... 74
    2.2.1 Waiving .................................................................................................................... 74
    2.2.2 The Four Relationships Hypothesis ...................................................................... 76

3 Networks of Trust .................................................................................................................... 85
  3.1 Conditions for Stability in Common Entrusted Party Networks .................................... 87
    3.1.1 The Main Source of Conflict.................................................................................. 87
    3.1.2 Two Ways to Rule Out (2)-Type Situations ......................................................... 89
    3.1.3 Uniform Discretion-Licencing Desires Networks in the Wild .............................. 96
    3.1.4 The Properties and Salience of Uniform Weighted Desires Networks 97
  3.2 Trust Networks Permitting Third-Party Disagreement .................................................. 103
    3.2.1 Why Might Conditions One to Five Be Met? ....................................................... 104
    3.2.2 Explaining Third-Party Disagreement in Photograph ........................................ 111
  3.3 ‘Harmony’ and the Reaction Problem ............................................................................. 113
3.3.1 The Admissible Grounds Hypothesis as it Applies to Entrusting Parties’ Preferences .................................................................115
3.3.2 The Hypothesis Applied to Discretion-Licencing Preferences ........119
3.3.3 Predicting ‘Harmony’ .................................................................................121
Conclusion ...........................................................................................................133

Part II Moral Psychology and the Moral Trust Network........................................135

4 Moral Judgement .................................................................................................136
  4.1 The Non-cognitive Attitudes Involved in Moral Judgements .......................136
  4.1.1 Moralists’ Voiceable Preferences ..............................................................136
  4.1.2 How Much I Need to Say about Moral Judgements .................................138
  4.1.3 The Negation Problem, and ‘Arrangements of Desires’ .........................140
  4.1.4 The Presumption of Dispositional Non-specificity ..................................147
  4.2 Disagreement in Matters of Basic Principle ..............................................149
Conclusion .............................................................................................................151

5 Moral Universality ...............................................................................................153
  5.1 Why Would the Moral Trust Network Exist? .............................................154
  5.2 Non-moral Precedents for Some Novel Features of the Moral Trust Network ................................................................................161
  5.3 Independent Evidence for the Existence of the Moral Trust Network ..........167
    5.3.1 Philippa Foot on Approval and Disapproval .........................................167
    5.3.2 Reason to Think that Entrusted Concern Relationships Enable Approval .............................................................................169
    5.3.3 That Something Like the Moral Trust Network is Needed to Explain Moral Approbation ..........................................................176
  5.4 Objections .......................................................................................................180
    5.4.1 Lack of Trust Owing to False Beliefs ...................................................181
    5.4.2 Moralists’ Trust is Implausibly Insensitive to Trustworthiness .............184
    5.4.3 Trust Between Antagonistic Agents .....................................................187
    5.4.4 Moral Judgement Without Altruistic Desires? .....................................188
Conclusion .............................................................................................................191

6 Moral Rules ..........................................................................................................192
  6.1 The Comparability Desideratum ..................................................................193
  6.2 ‘Norms’ and Simplicity .................................................................................195
    6.2.1 Fundamental and Derived Entitlements ...............................................196
    6.2.2 ‘Norm-Partitions’ and Related Terms .................................................198
    6.2.3 A Restatement of the Proposal, and Two Desiderata Derived From the Comparability Desideratum ...................................................200
  6.3 ‘Conforming to’ Norms .................................................................................202
    6.3.1 Three Precisifications Considered .......................................................202
    6.3.2 Solutions to the Problems Facing the Second Precisification ..............209
  6.4 Preferences for Conformity to Non-comprehensive Norm-Partitions ..........212
    6.4.1 Derived Voices and Norm-Indecision ....................................................212
    6.4.2 The Difficulty of Moral Reasoning, and its Practical Implications ..........216
    6.4.3 The Salience of ‘Norm-Principles’ ........................................................219
  6.5 Advantages of Simplicity ..............................................................................221
    6.5.1 How ‘Simplicity’ Aids the Comparative Method of Abuse Detection ..........221
6.5.2 Other Advantages ................................................................. 225
6.6 The Difficulty of Explaining Simplicity Desires .......................... 228
Conclusion .................................................................................. 230
7 Moral Motivation ....................................................................... 231
  7.1 Examples of Expressivists’ Commitment to Heft ....................... 232
  7.2 Impediments to Abandoning a Commitment to Heft .................. 236
  7.3 Problems with Heft and the Theory of Moral Reasoning it Supports .. 237
  7.4 Fiduciary Non-cognitivism’s Divorce from Heft ......................... 242
8 Moral Uncertainty ...................................................................... 243
  8.1 A Puzzle About Acting Under Moral Uncertainty ..................... 243
  8.2 Moral Uncertainty for Fiduciary Non-cognitivists ...................... 249
    8.2.1 A Refinement of My Account of Entrusted Parties’ Obligations and
    Entitlements ............................................................................. 249
    8.2.2 Confidence and Indecision in Matters of Basic Moral Principle ... 258
  8.3 A Fiduciary Non-cognitivist Solution to the Puzzle ................... 267
Conclusion .................................................................................. 269
9 Moral Forgetting ....................................................................... 270
  9.1 What Needs to Be Explained? .................................................. 271
    9.1.1 Behavioural and Phenomenological Symptoms of Forgetting in Moral
    Contexts ................................................................................. 272
    9.1.2 Forgetting Morally Relevant Non-normative Facts ............... 273
    9.1.3 Inverted-Commas Moral Forgetting .................................... 274
    9.1.4. Rhetorical Reports of Moral Forgetting .............................. 276
    9.1.5 Moral Forgetting Caused by the Forgetting of Moral Arguments ... 277
    9.1.6 Complexity ....................................................................... 279
    9.1.7 Memory Disorders ............................................................ 280
  9.2 Explanations .......................................................................... 281
    9.2.1 Commonsense Moral Scepticism and the Factivity of ‘Forget’ ...... 281
    9.2.2 Too Obvious to Forget ....................................................... 282
    9.2.3 Intuition ............................................................................ 282
    9.2.4 Speaker Subjectivism ......................................................... 284
    9.2.5 Too Important to Forget ...................................................... 288
    9.2.6 A ‘Companion in Absurdity’ .............................................. 289
    9.2.7 Traditional Non-cognitivism .............................................. 291
    9.2.8 Fiduciary Non-cognitivism ................................................ 295
Conclusion .................................................................................. 299

Appendices .................................................................................... 304
  Appendix A .................................................................................. 304
  Appendix B .................................................................................. 305
  Appendix C .................................................................................. 307
  Appendix D .................................................................................. 312
  Appendix E .................................................................................. 315
Bibliography .................................................................................. 319
Introduction

In this dissertation I present and defend a novel non-cognitivist theory of our moral thought and practices. Its most distinctive feature is its insistence that moral judgements depend, in hitherto unrecognized or under-recognized ways, on the existence of broadly ‘social’ relationships among those who make them. It boldly posits that all those who make moral judgements—all ‘moralists’—must be parties to a common network of trust relationships. To hold a moral judgement is necessarily to have a certain kind of desire-like internal mental state. But having the requisite internal mental state amounts to holding a moral judgement only if one is a member of the ‘moral trust network’. Moral judgements have the distinctive practical significance that they have, as states that can be expressed in order to justify, criticize, or demand actions, only because the moral trust network confers this significance on them.

The theory, which I call fiduciary non-cognitivism, has a number of significant virtues. Unlike most non-cognitivist theories, it avoids making our moral thought and practices appear very exotic, by characterizing them as more elaborate versions of certain familiar and humdrum kinds of deliberation and social interaction. It offers an account of what substantive moral reasoning involves that avoids reliance on the thesis that higher-order desires can (rationally) influence lower-order ones. It has a satisfactory psychology of moral uncertainty. Its explanation for the phenomenon of moral disagreement is superior to that of any rival non-cognitivist theory. The postulation of a potentially universe-spanning moral trust network is not as costly as it appears, and, as I argue, a non-cognitivist theory that doesn’t posit something like it will be unable to adequately account for some important phenomena.

Throughout this dissertation, unless I specify otherwise, whenever I talk of ‘rationality’ I have in mind rationality of a narrow and formal sort, which is broad
enough to encompass the more obvious and uncontroversial prescriptions of decision theory and formal (e.g. Bayesian) epistemology, as well as requirements of consistency among attitudes other than beliefs and preferences. It does not include ‘enkratic requirements’, requirements to respond correctly to one’s practical reasons, or requirements to do what ‘makes sense’, except insofar as these can be reduced to requirements of the ‘narrower’ sort I have outlined. I shall assume that the narrow rational requirements, or at least a large subset of them need not be construed as requirements of a richly normative sort, but rather may be understood as requirements of interpretive coherence or as entailed by a full description of the functional roles of attitudes.

In what follows I will say how I shall be using ‘non-cognitivism’ and related terms (0.1.1), briefly summarize some prominent existing versions of non-cognitivism (0.1.2), summarize the most common arguments for and against non-cognitivism (0.1.3), enumerate the metaethical desiderata that I shall be most concerned to meet in this work (0.2.1), say what fiduciary non-cognitivism is all about (0.2.2), give some indication of how fiduciary non-cognitivism meets the aforementioned desiderata (0.2.3), outline the scope of my project (0.2.4), make some comments in defence of my methodology (0.2.5), and summarize the contents of the thesis (0.2.6).

0.1 Non-cognitivism and its Place in Metaethics

0.1.1 Distinguishing Non-cognitivism from Cognitivism

I will first say something about what non-cognitivism is and its place in metaethical debate. Usage of the term ‘non-cognitivism’ varies; it is sometimes taken to refer to a family of theories specifically about moral (or normative) language, including emotivism, prescriptivism, and expressivism. However, I shall take moral non-cognitivism to be a theory or family of theories of moral thought and talk which deny that moral judgements are cognitive attitudes. Non-cognitivism is thus not purely a theory in linguistics or the philosophy of language: a complete non-cognitivist theory will be a theory of mind as well as language (and may make various negative claims.

---

1 Examples of the sorts of inconsistent combinations of attitudes I have in mind can be found in Derek Baker and Jack Woods, ‘How Expressivists Can and Should Explain Inconsistency’, *Ethics* 125 (2015): 401-409.

about ‘moral metaphysics’ too). By contrast, I shall regard expressivism as a purely linguistic theory, though one that is compatible with non-cognitivism.

But it is not a trivial task to say exactly what it means to affirm or deny that moral judgements are ‘cognitive attitudes’. Part of the problem is that moral thought and talk has many features that would, from an orthodox philosophical point of view, seem characteristically ‘realist’ and ‘cognitivist’. For instance, in folk usage moral claims can be believed, and asserted; they can be true, and can state facts. It is difficult to say what non-cognitivism is without characterizing it as the denial that moral thought and talk has these ‘realist’ or ‘cognitivist’ features—as the denial that moral judgements are beliefs, for instance. And yet we should, at least to a certain extent, avoid characterizing non-cognitivism in this way. This is because many non-cognitivists have adopted positions in the philosophy of language that purportedly allow them to avoid denying morality’s ‘cognitivist’ appearances. They have embraced deflationary, minimalist construals of folk terms like true, believe, assert, proposition, and so on, that purportedly allow them to vindicate talk of moral truth, belief, assertion, and so on, while maintaining that moral judgements are desire-like mental states. The same considerations seemingly favour, and the same strategy would facilitate, the vindication of talk of moral claims describing and representing states of affairs. If a minimalist construal of these terms is correct, cognitivists and non-cognitivists alike hold that moral judgements and/or moral claims are truth-apt, doxastic, assertoric, propositional, descriptive, and so on. But this agreement makes it hard to give an account of the distinction between cognitivism and non-cognitivism.\(^3\)

As there is no generally accepted account of this distinction, I shall impose my own favoured account here. I think that there is a subset of ‘cognitivist-sounding’ folk terms of which the non-cognitivist does not need to give a minimalist construal. The subset includes the terms describe and information, and their cognates.\(^4\) These terms can certainly be used felicitously in connection with moral claims, as in the following:

---


4 The subset also includes the terms forget and remember, for reasons that I give in Chapter 9. It may include represent, depict, and portray. However, it is difficult to construct examples, along the lines of (1*) and (2*) below, with which to gauge the desirability of ‘deflating’ these terms. Such examples will often be confounded by an abundance of available interpretations. E.g. ‘Practical Ethics portrays some instances of infanticide as permissible’ does not sound very odd, but this may just be because we can interpret it as meaning that the infanticides are presented in such a way as to encourage us to think them permissible. The non-cognitivist
(1) In Cliff’s Notes on Macbeth, Macbeth is described as a good man turned bad by ambition.\(^5\)

(2) I haven’t read Macbeth. But Cliff’s Notes on Macbeth contains the information that Macbeth was a good man turned bad by ambition.

But in some contexts their use in connection with moral claims seems careless or odd:

(1\(^*\)) In Practical Ethics, failure to kill babies in some situations is described as wrong.

(2\(^*\)) I haven’t thought very hard about infanticide. But Practical Ethics contains the information that failing to kill babies in some situations is wrong.

My intuition about (1\(^*\)) is very much what we should expect, given a philosophically orthodox understanding of ‘describe’, if non-cognitivism were true. It seems like a mistake to talk of describing infanticide as morally required, because describing X as such-and-such is saying something about X—it is ruling out some ways X might be. But saying that X is permissible isn’t saying anything about X, and doesn’t rule out any other ways that X might be.\(^6\) My intuition about (2\(^*\)) is somewhat different. By calling the claims about infanticide in Practical Ethics ‘information’, (2\(^*\)) places them in the same class as the mundane claims of an encyclopedia, a newspaper, or Cliff’s Notes, as things we might just accept without feeling any particular need to scrutinize them. It thus seems odd in much the same way that moral testimony seems odd.\(^7\)

The following explanation is available as an explanation for the difference between (1)-(2) and (1\(^*\))-(2\(^*\)): we are inclined to interpret the first, but not the second, as, in part, claims about the description of non-moral facts that are presumed to be

seemingly does not need to be a deflationist about ‘portray’ to vindicate this usage. But there is perhaps a sense of ‘portray as’ that is more like ‘describe as’ whose use in moral contexts could not be vindicated by the non-cognitivist without deflation. The example does not help to establish whether this usage deserves to be vindicated.

\(^5\) I don’t know whether Cliff’s Notes says any such thing. These examples are based on an example from James Dreier (‘Creeping Minimalism’, 29).


\(^7\) There is a sizeable literature on why moral judgement-formation on the basis of testimony seems strange or defective, and whether this appearance is accurate, which largely proceeds on cognitivist assumptions (e.g. Sarah McGrath, ‘Scepticism About Moral Expertise as a Puzzle for Moral Realism’, Journal of Philosophy, 108, no. 3 (2011): 111-137; Sarah McGrath, ‘The Puzzle of Pure Moral Deference’, Philosophical Perspectives 23 (2009): 321-44; Alison Hills, ‘Moral Testimony and Moral Epistemology’, Ethics 120 (2009): 94-127; Robert Hopkins, ‘What is Wrong With Moral Testimony?’, Philosophy and Phenomenological Research 74, no. 3 (2007): 611-634; Karen Jones, ‘Secondhand Moral Knowledge’, The Journal of Philosophy 96, no. 2 (1999): 55-78). Assuming the non-cognitivist does not want to go out of her way to make room for moral testimony, it is natural for her to explain these appearances by simply denying that coming to hold moral judgements on the basis of moral testimony, understood as a potentially rational way of acquiring moral judgements, is possible. It is far from obvious that there is any rational process by which we can communicate desire-like attitudes to one another.
morally relevant, or about non-moral information that is conventionally presumed to be morally relevant.\footnote{The idea that moral utterances communicate non-normative information is a familiar one, and we have ample reason to think that some version of it is true (see R. M. Hare, *The Language of Morals* (Oxford: Oxford University Press, 1952), 111-23 on the ‘secondary’ descriptive meaning of evaluative utterances). I am suggesting that (1) and (2) are felicitous because we understand that it is the non-normative implications of *Cliff’s Notes*’ moral claims about Macbeth that they are characterizing as descriptive or as information.} For instance, we are inclined to interpret (1) as implying ‘*Cliff’s Notes on Macbeth* describes Macbeth as having non-normative properties that we all think would make him a good man turned bad by ambition.’ By contrast, we are not inclined to interpret (1*) merely as an implicit claim about how *Practical Ethics* describes the non-normative properties of infanticide in the relevant cases. No claim that infanticide is permissible is likely to be based purely on the principles of conventional morality, plus some interesting non-normative claims. If this explanation is right, and if it is only the availability of the elliptical explanation I have described that explains the felicity of (1)-(2), then there may well be no genuine talk of moral description or moral information that warrants vindication.

I shall assume that this is correct, and that the non-cognitivist’s deflationary program stops short of positing a minimalist semantics for these terms. I shall therefore characterize moral non-cognitivism as the metaethical theory that moral judgements are non-descriptive, and that there is no moral information. I shall characterize cognitive attitudes as attitudes that are or purport to be information-bearing, and doxastic attitudes as the subset of these consisting of cognitive beliefs and degrees of belief. I shall generally assume that non-cognitivists take moral judgements to be non-cognitive attitudes of some kind, but I won’t treat this as a necessary feature of non-cognitivism.\footnote{It may be possible for a theorist to be recognizably non-cognitivist while denying that moral attitudes exist and that moral utterances purport to express them. I have in mind a prescriptivist who holds that moral utterances are commands, and denies that there is any unified class of attitudes that they conventionally express.}

I should acknowledge that the position I am adopting here is not wholly satisfactory, because the intuitions that (1*)-(2*) are infelicitous are subtle and fragile, and some readers may lack them altogether. I would be hesitant to adopt it if an alternative position were well established in the literature.

Throughout this work, with the exception of Chapter 9, I will try to avoid saying anything whose truth either requires or is incompatible with minimalism about terms other than ‘describe’ and ‘information’. In accordance with metaethical tradition, I will
use ‘moral judgement’ as a neutral term for the things that the folk ordinarily call ‘moral beliefs’ which implies nothing about their underlying psychological nature.

0.1.2 Prominent Contemporary Non-cognitivist Theories

Contemporary versions of non-cognitivism are almost invariably conjoined with expressivist theories of moral language. Important pre-expressivist theories, such as the emotivist theories of A. J. Ayer and Charles Stevenson, and the prescriptivism of R. M. Hare, can (according to Mark Schroeder) be thought of as speech act theories. We can, with some interpretive licence, construe them as holding that the speech acts that declarative moral sentences are best-suited to perform constitute the primary meanings of those sentences. Expressivists, by contrast, are not reliant on hypotheses about the characteristic illocutionary force of moral sentences for their theories of meaning.10

Moral expressivism is roughly the thesis that (i) declarative moral sentences express non-cognitive attitudes in the same way in which declarative descriptive sentences express cognitive beliefs, and (ii) the meaning of a moral sentence consists in its expressing the non-cognitive attitude it expresses.11

The most prominent contemporary non-cognitivist theories are those of Simon Blackburn and Allan Gibbard. The central idea of expressivism was largely developed in Blackburn’s writings. Blackburn is a leading proponent of the ‘quasi-realist’ project of accommodating the ‘realist’ appearances of morality, in part by means of the minimalist program that I discussed in subsection 0.1.1. He has also made several somewhat piecemeal attempts at providing an expressivist account of normative language.12 These semantic proposals are historically important, but few if any expressivists now accept them.13 Blackburn seems to have focused more on the large meta-philosophical issues raised by expressivism, and on prosecuting the case against cognitivism, than on providing a detailed non-cognitivist account of moral thought and talk.

Allan Gibbard’s non-cognitivist theorizing has been more detailed and systematic. In Wise Choices, Apt Feelings, Gibbard presents ‘a theory of the meaning of moral

10 Mark Schroeder, Noncognitivism in Ethics (London: Routledge, 2010), 74.
11 Schroeder, Noncognitivism in Ethics, 72-3.
13 Mark Schroeder, Being For: Evaluating the Semantic Program of Expressivism (Oxford: Oxford University Press, 2008), 10n.3; Noncognitivism in Ethics, 122.
terms’ that ‘rests on a psychology of norms’. His ‘norm-expressivist’ theory of meaning is, at its broadest, a theory about claims of rationality, or of what actions and mental states it ‘makes sense’ to do or have. Gibbard held that such judgements are non-cognitive attitudes of norm acceptance (combined, perhaps, with relevant descriptive beliefs). To say that φ is irrational on this view is, roughly, to express one’s acceptance of norms forbidding φ. If one accepts a set of norms N one has some tendency to conform to N. Gibbard characterizes norm acceptance in part by contrasting it with other motivationally efficacious relations we can stand in toward norms, such as being ‘in the grip’ of social norms, or as having ‘internalized’ norms; and by positing that it is the mental state that struggles against our appetites in cases of weakness of will, that we are disposed to avow in group deliberation, and that is capable of being shaped through such deliberation. Moral judgements are a subset of judgements of rationality according to Gibbard, and so fall within the scope of his theory of rational judgement. They are, roughly, attitudes of accepting norms concerning when to feel guilt and resentment.

To say that an action φ is irrational is to express a norm-acceptance and belief state that ‘rules out’ every pair consisting of a system of norms permitting some type of action and a world in which φ is an action of that type. A system of norms is a set of prescriptions dividing all actions into forbidden, permitted, and required subsets. Complex sentences in which atomic moral sentences are embedded work in a similar fashion. For instance (borrowing an example of Gibbard’s), to say, ‘Whenever Antony does anything irrational, he sticks to his purpose stubbornly’ is roughly to express a mental state that rules out every pair consisting of a system of norms forbidding something and a world in which Antony does that thing without sticking to his purpose stubbornly.

0.1.3 Standard Arguments For and Against Non-cognitivism

The main reason to posit that moral judgements are desire-like rather than information-bearing mental states is that this allows us to explain the intuition that there is a necessary connection between judging that one morally ought to φ and having some

motivation to φ (‘motivational’ or ‘judgement internalism’), in a manner consistent with the Humean Theory of Motivation, which says that cognitive attitudes alone cannot rationally motivate action.¹⁸

Many non-cognitivists also hold that their moral realist rivals face a dilemma which they themselves avoid. Moral realists claim that moral judgements are cognitive attitudes, that to say that an action is morally wrong, permissible, good etc. is to purport to give information about it, and that some non-trivial moral claims are true. Moral realists must take moral predicates to refer to either natural or non-natural properties. If moral properties are natural properties, it becomes unclear why people who differ greatly in the actions to which they apply and refuse to apply terms like ‘morally wrong’ (or as I shall vaguely put it, who ‘apply these terms to very different things’) do not simply mean different things by those terms. Yet people who, in their sincere assertions, apply these terms to very different things very often do seem to ‘mean the same thing’ by them, and thus to be in disagreement with each other when they apply them to different things. Explaining this apparent disagreement is thus alleged to be a challenge for the realist who takes moral predicates to refer to natural properties.¹⁹

Those realists who hold that moral properties are non-natural properties, on the other hand, face a serious challenge when it comes to explaining the ‘supervenience of the moral on the non-moral’. I shall characterize the datum of moral supervenience psychologically: it is the thesis that one necessarily errs if one judges that one action is wrong, good, etc., and that another is not, while judging that these actions do not differ in their natural properties. To predict supervenience, the realist must posit that her hypothesized moral properties are such that, necessarily, two actions only differ in their

---


¹⁹ Blackburn, Spreading the Word, 168. This objection is thought to be more compelling if (i) each speaker is from a different speech community, (ii) each applies the relevant word (e.g. ‘wrong’) to the same sorts of things that others in her speech community tend to apply it to, and (iii) the word has the same sort of role in deliberative and critical discourse in each speech community as it has for us. It is the common practical significances of the word that is supposed to assure us that the speakers ‘mean the same thing’ by it. See Hare, Language of Morals, 148-9; Terence Horgan and Mark Timmons, ‘Troubles for New Wave Moral Semantics: The “Open Question Argument” Revived’, Philosophical Papers 21, no. 3 (1992): 153-175. For a readable overview of the topic, see Tristram McPherson, ‘Semantic Challenges to Normative Realism’, Philosophy Compass 8, no. 2 (2013): 126-136.
moral properties if they differ in their natural properties. This necessary relationship demands an explanation, and it is not clear that the non-naturalist realist can supply one.\(^{20}\)

How does non-cognitivism avoid these problems? Non-cognitivists claim to be able to explain moral disagreement as a clash of non-cognitive attitudes. Diversity in the normative content of moral judgements among different people and different cultures is no obstacle to this sort of explanation,\(^ {21}\) save in extreme cases. Non-cognitivists have more options for explaining supervenience than non-naturalist realists do, because on their understanding moral judgements do no information-bearing duty. Non-cognitivists can posit that moral judgements are subject to a ‘supervenience constraint’ because this serves some broadly practical function.\(^ {22}\) On cognitivist assumptions, this postulation would be implausible because, if (i) moral judgements purported to be moral-information-bearing states tracking moral properties and (ii) moral properties did not supervene on natural properties, a ‘supervenience constraint’ would interfere with moral judgements’ purported information-bearing role. But the truth of this conditional is no embarrassment to non-cognitivists because they deny (i). So non-cognitivists avoid both horns of the dilemma facing realists.

These are the oft-cited advantages of non-cognitivism. The theory also has weighty disadvantages. The weightiest is the so-called Frege-Geach problem. This is a problem for non-cognitivist theories of moral language. A non-cognitivist needs to supply such a theory: to give an account of what sentences containing moral terms mean, or of their linguistic purpose, that explains the way in which they are used. In ordinary use, some such sentences imply or are inconsistent with others, so the non-cognitivist’s theory of moral language must explain this. The most significant attempts to explain this feature of moral language have been versions of expressivism.\(^ {23}\) If two


\(^{21}\) See e.g. Blackburn, *Spreading the Word*, 168.

\(^{22}\) Blackburn, *Spreading the Word*, 186.

\(^{23}\) Some non-cognitivists (broadly construed) advance expressivism as a metasemantic rather than a semantic thesis, and give a non-expressivist (e.g. truth-conditional) semantics for moral sentences. See Michael Ridge, *Impassioned Belief* (Oxford: Oxford University Press, 2014), esp. 8-9, 103ff.
sentences are inconsistent, the expressivist holds that this is owing, fundamentally, not
to the fact that they cannot both be true, but to the fact that it is rationally inconsistent to
simultaneously hold the attitudes they express. If one sentence ‘P’ implies another
sentence ‘Q’, this is said to be because it is irrational to have the attitude expressed by
‘P’ and the attitude expressed by the negation of ‘Q’, and because, in the right
circumstances, it is rationally permissible or mandatory to ‘transition’ from holding the
attitude expressed by ‘P’ to holding the attitude expressed by ‘Q’. (Something similar
holds for relationships of inconsistency and implication among different sets of
sentences.)

Expressivist theories face no formidable and immediately apparent problems in
accounting for the meanings of ‘atomic’ moral sentences, i.e. indicative sentences
consisting merely of a subject and a moral predicate. But more complex sentences
involving disjunction and negation, and ‘mixed’ sentences constructed from atomic
moral sentences and non-moral sentences, are another matter.

The Frege-Geach problem is roughly the problem for non-cognitivists of giving a
constructive account of the meaning of moral sentences, both atomic and complex, that
explains the inconsistency of intuitively inconsistent sets of sentences containing moral
terms, and the validity of intuitively valid arguments containing moral terms. On most
accounts, expressivists have thus far failed to provide a solution to the Frege-Geach
problem that is not prohibitively costly in some respects.24 And, though the Frege-
Geach problem has engrossed the attention of non-cognitivists who have turned their
attention to semantics, it is only one facet of the problem of providing an adequate
expressivist semantics.25

24 Unwin’s ‘negation problem’ seems to refute the expressivist theories of Blackburn and
Gibbard. Nicholas Unwin, ‘Quasi-Realism, Negation, and the Frege-Geach Problem’, The
Philosophical Quarterly 49, no. 196 (1999): esp. 341-2; ‘Norms and Negation: A Problem for
levelled a similar and similarly compelling objection against Michael Ridge’s explanation for
normative disagreement, which Ridge has conceded. James Dreier, ‘Truth and Disagreement in
Impassioned Belief’, Analysis 75, no. 3 (2015): 457-8; Michael Ridge, ‘Replies to Critics’, same
volume, 479. In Being For, Mark Schroeder develops an expressivist theory that solves the
‘negation problem’, but which either cannot explain the logical properties of sentences with
embedded atomic normative and non-normative sentences, or is incompatible with a plausible
semantics for descriptive sentences. (Schroeder is of the opinion that no other expressivist
theory can do better, and concludes that he has shown sufficient reason to reject expressivism.
Being For, 179.)

25 See Mark Schroeder’s Being For for an evaluation of the prospects for an expressivist
account, both of normative logic, and also of tense, modality, and quantification, especially as
they apply to descriptive sentences.
0.2 The Project

0.2.1 Metaethical Desiderata that are Relevant to the Project

I’ll now describe the explanatory challenges that I will be concerned with, and briefly comment on how well existing theories meet them, giving particular attention to non-cognitivist theories.

Disagreement. First, a moral psychology must explain why two people seem to disagree when they hold moral judgements that we would pretheoretically take to be inconsistent. And the predicted intuition of disagreement should be of the right sort. The intuition of disagreement elicited by such cases is not weak, impressionistic, or questionable, but plain and obvious, such as that elicited by ordinary cases involving inconsistent descriptive beliefs.26 A moral psychology should predict this. Among cognitivists, moral naturalists arguably have difficulties meeting this desideratum, as I noted above, and so do various sorts of relativist. Non-cognitivists certainly face challenges here. There are cases of ‘clashing’ (e.g. mutually unsatisfiable) non-cognitive attitudes that seem clearly to be disagreements, but there are also plenty that don’t. Non-cognitivists need to supply a plausible theory of non-cognitive disagreement that will convince us that a clash of non-cognitive attitudes which, according to their theories, constitutes a moral disagreement would have the correct appearance of disagreement. A few non-cognitivists have tried to devise theories of roughly this sort, though none successfully.27

Justification, criticism, and demands for grounds. Second, moral statements can be used to defend and criticize actions, and an account of our moral psychology and practices must explain this. There are difficulties here for various views in metaethics, including non-cognitivism. It is not clear why letting someone know that you have a desire-like attitude should have the upshot that we expect justificatory or condemnatory utterances to have. Suppose that someone objects to my having φ-ed, and I say in

26 There are admittedly clashes of descriptive belief that seem ambiguous. I discuss a few in 1.1.6.

27 Gibbard defends a theory of disagreement in plan in Thinking How to Live using an elaborate argument of a sort that Michael Ridge characterizes as ‘transcendental’. Allan Gibbard, Thinking How to Live, (Cambridge, Mass.: Harvard University Press, 2003), Chapter 14; Ridge, Impassioned Belief, 181. Ridge offers a summary and critique of Gibbard’s theory that I don’t think I can improve upon. He argues that Gibbard’s theory is unclear on crucial points and fails to explain all of the intuitive data (Ridge, Impassioned Belief, 179-85). I shall defer to Ridge on these points. I offer counterexamples to a theory of non-cognitive disagreement resembling Charles Stevenson’s in 1.1.3, and to Michael Ridge’s theory in 1.1.5.
defence of my action that φ-ing was the right thing to do. The expressivist theory that Gibbard defends in *Thinking How to Live* might characterize this as expressing a plan to φ in the circumstances in which I acted. T. M. Scanlon wonders, ‘why should this carry any weight with the person who has complained? It is already obvious that I plan to behave in these ways. That is what the person is objecting to.’ The problem, or some version of it, remains, at least according to Scanlon, if we say that normative statements express ‘acceptance of norms’, imperatives, or approval rather than plans.

It seems desirable that an explanation of our justifying and criticizing practices should account for certain further phenomena. When someone uses a moral claim to condemn or justify an action by saying that it is or isn’t wrong, it is appropriate to respond by demanding that she say why it is wrong or not wrong. The justifying or condemning party then normally seems obliged to give such an explanation or withdraw her justification or criticism. There seem to be ways in which such an explanation can be more or less ‘successful’, and its ‘success’ can affect how seriously the attempted justification or criticism is taken. It would seem relatively easy for cognitivists to account for these further phenomena, assuming that they can account for the use of moral claims to justify and criticize, for it is uncontroversial that there are descriptive assertions for which explanations can be demanded, and that they can be explained more or less successfully. But it is not clear how non-cognitivists can account for it.

*Moral uncertainty*. Third, moral judgements can be held with varying degrees of certainty, even by an agent who is perfectly certain about all empirical facts that she takes to be potentially morally relevant. Theories of moral psychology must explain this phenomenon. This is a problem for non-cognitivism because non-cognitive attitudes apparently cannot be held with differing degrees of certainty. According to one recent commentator, non-cognitivists have so far failed to supply a wholly satisfactory account of normative uncertainty. Cognitivists have no trouble saying what moral uncertainty is. But an account of moral uncertainty must make the right predictions about the consequences of moral uncertainty for rational action, and I will argue that some cognitivist theories struggle here. These theories either fail to explain why it is ever

---

29 Scanlon, *Being Realistic*, 59. I am not convinced that the problem remains if we say that normative statements express states of approval and disapproval. But I think that non-cognitivists should avoid postulating this until they can offer a convincing theory of what approval is.
rational to act on uncertain judgements of basic moral principle, or they have counterintuitive implications concerning how people will act under moral uncertainty.

*Moral reasoning and changes of motivation.* Fourth, a moral psychology must explain how substantive moral reasoning can produce changes of moral judgement and hence changes in moral motivation. This is a challenge for cognitivists who, like Michael Smith, maintain that the motivational changes associated with changes of basic moral judgement (i.e. judgements of basic principle) are principally changes of basic or non-instrumental desire.\(^{31}\) Non-cognitivists typically characterize substantive moral reasoning as a mental activity aimed at making one’s non-cognitive moral attitudes better satisfy some subset of one’s higher-order non-cognitive attitudes (e.g. one’s desires concerning which non-cognitive moral attitudes one should have).\(^{32}\) This is an imperfectly satisfactory account because no very satisfactory explanation can be given as to how the changes in one’s first-order moral attitudes are brought about. The most natural story is as follows: it is rationally incoherent to have higher-order non-cognitive attitudes that are not satisfied, we have some sort of mental or neurological faculty for ridding ourselves of incoherence, and this faculty is what causes the changes of mind that are not uncommonly the upshot of moral reasoning. But it is far from clear that there is necessarily any incoherence involved in having frustrated higher-order non-cognitive attitudes.\(^{33}\)

*The existence of morality.* Fifth, it is desirable that an account of our moral psychology and practices should be such that the existence of our moral psychologies and practices, as they are characterized by the account, should not be too empirically surprising.\(^{34}\) An account should avoid construing moral thought and practices as involving behaviours, mental states, or mental processes that both (i) do not manifestly occur, and (ii) could not be well explained given our existing knowledge in the social sciences, psychology, evolutionary biology, and so on, if they did occur.

---

\(^{31}\) See Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994), 72-76 on the underived character of moral desires. For Smith’s solution to the problem, see Chapter 5, esp. 159-161.

\(^{32}\) I provide textual support for this point in Chapter 7.

\(^{33}\) For a very brief survey of views on the rationality of having unsatisfied second-order preferences, see Alejandro Perez Carballo, ‘Rationality and Second-Order Preferences’, *Nous* (‘Early view’ version) (2016): 2-3. I discuss Carballo’s own explanation of how coherence can require the satisfaction of second-order preferences in Chapter 7.

\(^{34}\) Of course, this is not a desideratum for the metaethicist who holds that moral thought and practices do not exist. Such a theorist is instead obliged to provide an error theory to explain why the folk have been misapplying moral-psychological (etc.) concepts.
Non-cognitivist views, it seems to me, typically meet this sixth desideratum poorly. They characterize moral thought and talk as quite different from kinds of thought and talk of which we have a better understanding, and postulate *sui generis* attitudes rather promiscuously (or tacitly rely on such postulations).\(^35\) Assuming that moral thought and talk are shaped fundamentally through evolutionary processes, this sort of characterization gives evolution a lot more work to do, so to speak. And non-cognitivists have provided little independent evidence that evolution has done this work. So, for all non-cognitivists have said, the existence of morality in its familiar form is empirically more surprising given non-cognitivism than it would be given some rival theories.

Allan Gibbard has, I acknowledge, gone to considerable lengths to provide an evolutionary explanation for the existence of normative—including moral—thought and talk construed in non-cognitivist and norm-expressivist fashion. But I don’t think he says enough to ameliorate non-cognitivism’s comparative shortcomings regarding this desideratum. According to Gibbard, our hunting and gathering ancestors faced many coordination problems, especially bargaining problems.\(^36\) An inability to coordinate in this environment tended to diminish an individual’s reproductive success. The chief biological function of normative thought and talk is to help us solve these problems—to coordinate our actions and emotions.\(^37\) Normative judgements motivate, and normative discussion tends toward consensus, so normative practices tended to give our ancestors shared ends. Coordination between people is more probable the more their ends align.\(^38\) So our normative psychologies promoted our ancestors’ reproductive success.

But however plausible Gibbard’s evolutionary story may be,\(^39\) it doesn’t seem to provide much support for norm-expressivism in particular. It seems to support any

\(^{35}\) On this last point, see Mark Schroeder’s repeated complaints in Chapter 3 of his *Being For*.

\(^{36}\) Gibbard, *Wise Choices*, 67. Two parties face a bargaining problem if it would benefit both of them to reach any of a range of agreements and, within this range, the agreement that would most benefit one party differs from that which would most benefit the other.


\(^{38}\) Gibbard, *Wise Choices*, 72-3. Gibbard does not explicitly talk about normative discussion’s tendency to give agents shared ends, but I think that this must be an important part of his story.

\(^{39}\) My inexpert objection to the account, as it applies to morality, is this. (1) It is not very plausible that our practice of engaging in moral *argument* would have evolved because of its tendency to bring about consensus, for moral argument is notoriously intractable. Striking bargains is much easier than reaching moral agreement through argument, because moral judgments resist pragmatic compromise. (2) The phenomenon of non-rational moral *influence*—e.g. the tendency of people’s moral judgements to ‘rub off’ on each other—might admittedly have evolved because of its tendency to produce shared ends. (‘Mutual influence’ is one of
naturalistic metaethical theory that predicts a fairly robust (though perhaps somewhat contingent) connection between moral judgement and motivation, that predicts a tendency toward consensus in moral discourse, that otherwise adequately explains the obvious facts about the subject matter of metaethics, and that is not premised on its own special (e.g. contractarian) story about why our moral practices exist. If I am right about this, Gibbard’s explanation for the existence of moral thought and talk is incomplete, for it does not explain the particular non-cognitivist form that morality takes. It thus does not make the existence of morality on a non-cognitivist construal much less surprising.

Ubiquity of moral thought and talk. Most accounts of our moral practices meet my sixth desideratum effortlessly, but it poses a serious challenge for my theory. The desideratum concerns the fact that the capacity to participate in moral thought and talk is pervasive among humans. People almost everywhere seem to be capable of making moral judgements, and when they do so, they seem to be engaging in the same practice. Not just the same broad kind of practice; rather, they are all, in a sense, making moves in the same game. A person who, without explicit qualification, condemns abortion on what we would intuitively regard as moral grounds is condemning abortion all over the world, not just abortion performed by members of her own community. She disagrees with and opposes anyone in the world who judges abortion to be permissible (on intuitively moral grounds) and sides with anyone else who condemns it. Bracketing differences of language, two strangers from foreign cultures can engage in moral discourse with each other, whether they or members of their respective cultures have encountered each other before or not.

A theory of moral thought and talk must explain this. The problem of meeting this desideratum is similar to the alleged problem, discussed above, for moral naturalists of explaining why people who apply ‘wrong’ to very different things seem to disagree instead of merely seeming to mean different things by the word. However, the present problem is in at least one respect more general, for it does not arise only in cases in morality’s consensus-creating mechanisms on Gibbard’s account (Wise Choices, 73.). But roughly the same result could have been achieved by the evolution of a simple propensity to desire the same states of affairs that those around you desire (whether through a ‘rubbing-off’-type process, producing lasting changes in desire, or something more like empathy). It is thus implausible to posit that the elaborate edifice of morality evolved merely to serve this purpose.

If predicting this were desirable, a cognitivist could do so fairly easily by pointing out (1) our willingness to accept testimony in descriptive matters, and (2) the tendency of arguments and evidence to reduce the diversity of opinions in certain descriptive domains.
which the parties (and their speech communities) apparently apply moral terms to quite
different things.

There are possible metaethical views that struggle to explain how people from
different isolated communities could be morally agreeing, or participating in the same
kind of practices, even when they and their communities apply a moral predicate to
apparently very similar actions, and use the word with the same intended practical
upshot. I don’t think cognitivists in general face a problem here, as they can plausibly
maintain that these people are all participants in the same ‘game’, viz., that of
describing how things are morally.

For purposes of illustration, a theory that would face a challenge here is one that
explained morality as arising from an actual social contract (rather than the more
commonly-appealed-to hypothetical sort), and that characterized moral judgements as
judgements about the terms of that contract.41 Such a theory would struggle to explain
why people in diverse and non-interacting communities who used what were apparently
moral terms in apparently the same way (applied to the same actions, and having the
same justificatory force etc.) would be talking about the same social contract in doing
so. How could they have ‘signed up’ to the same social contract when their
communities were isolated from each other?42 Fiduciary non-cognitivism also struggles
to meet this desideratum, for similar reasons.

Rule-like character of moral requirements. There is a tendency among makers of
moral judgements to think that particular moral requirements (and permissions etc.) can
always be explained by general requirements. If someone says a particular action is
wrong, for instance, it is always reasonable to ask why. And a paradigmatically
appropriate answer will be a claim, for a somewhat simple natural property F, that all
actions that are F are wrong, and that the action in question is F. In other words, there is
a presumption that moral requirements will take the form of general rules. This

41 Jean Hampton puts the case for a view broadly resembling this in her *Hobbes and the Social
Contract Tradition* (Cambridge: Cambridge University Press, 1986). Cf. also Gilbert Harman,
Southwood for these references.
42 Of course they might have ‘signed up’ to different contracts, and yet be talking about the
same contract—party A might be talking about party B’s contract, and B might also be talking
about B’s contract. But this creates a puzzle as to how A could be using moral words with any
kind of justificatory or condemnatory ‘force’, for it would presumably have been only qua
member of her own community’s contract that A could use these words with this kind of force.
Plausibly, insofar as A was describing B’s social contract, she would be talking in a quasi-
anthropological fashion, not talking qua moralist.
presumption tends to survive the realization that there are few interesting, general, precisely specified, and non-gerrymandered moral rules that avoid making unpalatable prescriptions in some circumstances. The seventh desideratum on my list is that of explaining this tendency to think of moral requirements as consisting of general rules.

This phenomenon does not seem to pose a problem for cognitivists, who can explain it as an instance of a more general human tendency to favour simple theories over complex ones in reasoning involving cognitive attitudes. We expect morality to be explained by simple laws just as we expect the world to be explained by simple laws. The desideratum is potentially more challenging for non-cognitivists. It is not clear, for instance, that humans have any particular tendency to form desires or preferences that are simple and ‘rule-like’ in character (it is in fact not clear to me what sort of desire would correspond to a judgement of moral requirement, where the requirement was of a ‘rule-like’ character). This desideratum will pose a challenge for my non-cognitivist theory of moral psychology.

*Moral forgetting*. The final two desiderata concern intuitive appearances that have received comparatively little attention from metaethicists. The eighth concerns our intuitions about moral forgetting.\(^{43}\) Cases in which people are presented as having lost moral beliefs through forgetting are apt to puzzle us, and an account of moral psychology should account for this. Establishing which views meet this desideratum and which do not is a large task, which I undertake later in this dissertation. But prima facie, non-cognitivist theories tend to meet it well, for they typically posit that moral judgements are attitudes of a sort that, at least intuitively, cannot be lost through forgetting. Cognitivist views will explain it better if they posit that moral facts are of an unusually ‘unforgettable’ sort, perhaps because they are intuitively obvious or discoverable through easy introspection.

*Moral approval and disapproval*. Ninth, it is desirable that a metaethical theory should be compatible with a unified account of the phenomena of approval and disapproval. As Philippa Foot argued, approval and disapproval on non-moral grounds seems to be impossible in the absence of certain type of social setting. For instance, a parent can disapprove of her daughter’s engagement on the grounds that the prospective husband is not rich enough, or not well enough connected, but a stranger cannot

\(^{43}\) The classic discussion of this topic is Gilbert Ryle’s ‘On Forgetting the Difference Between Right and Wrong’, in *Essays in Moral Philosophy*, ed. A. I. Melden ed. (Seattle: University of Washington Press, 1958), 147-59.
disapprove of the engagement on the same grounds. Foot proposes that this is because non-moral approval and disapproval require the existence of conventions entitling the approving or disapproving party to be listened to in the decisions of the approved or disapproved party. Convention, for instance, entitles a young woman’s parent, but not a stranger, to a voice in her marriage choices. It is prima facie desirable we should give a unified account of the conditions for approval and disapproval. So if we are persuaded by Foot’s intuitive data, as I think we should be, then it is desirable that we should say, as Foot does, that even a moral disapprover must enjoy an entitlement to be listened to in the relevant decisions of those of whose actions she disapproves. Meeting this prima facie desideratum would mean explaining how such entitlements could exist. (I call this a ‘prima facie desideratum’ because it may turn out that there is no way to give a unified account of the conditions for approval and disapproval that is not prohibitively costly.)

This concludes my enumeration of the metaethical desiderata that are relevant to my present project. Fiduciary non-cognitivism deserves the reader’s attention because it meets them tolerably well—much better, on the whole, than any other extant non-cognitivist theory—and at no detriment to its explanation for moral supervenience or the necessary connection between moral judgement and motivation. It is, moreover, a unified theory, which relies on a fairly small number of novel postulations, quite a few of which turn out to be independently plausible on scrutiny. The theory is incomplete, however, and has a few worrisome and distinctive commitments. (I mention most of the theory’s problems in 0.2.3 and 0.2.4 below.) I will now say a bit more about the nature of the theory defended in this dissertation.

0.2.2 The Central Idea of Fiduciary Non-cognitivism

If two people seem to disagree in virtue of having clashing non-cognitive attitudes, the appearance of disagreement will generally be weak and fragile unless there is a ‘matter to be settled’ between them. The ‘matter to be settled’ cannot be a mere conflict, in which each party regards the other merely as a threat to her interests. Rather, one (or each) party must have expectations, of a broadly normative sort, about how the other will behave, and the other must somehow acknowledge and aspire to meet these expectations. So I shall argue in this thesis.

I hypothesize that these normative expectations, if they are justified, are explained by relationships of trust between the parties. Specifically, one party can be in ‘robust’
non-cognitive disagreement with another—i.e. have a clash of non-cognitive attitudes with the other of a sort that elicits a robust intuition of disagreement—only if she shares with the other a certain kind of trust relationship. I call relationships of this sort ‘entrusted concern relationships’. In an entrusted concern relationship, one party entrusts to another a concern to accommodate those of her preferences that meet certain criteria. The latter party violates her trust if he resists her preferences, unless he does so in conformity to preferences of his own that meet certain criteria. A clash among preferences that meet the relevant criteria for each party amounts to a robust non-cognitive disagreement. The criteria are determined by the ‘terms’ of their trust relationship.

Judgements are capable of ‘robustly’ disagreeing; a clash of judgements elicits a firm intuition of disagreement.\(^44\) This is a constraint on what could count as a moral judgement. A given attitude J could not be a moral judgement unless there were some other attitude K such that having J necessarily puts a person in ‘robust’ disagreement with any other person who has K. A non-cognitivist who hopes to meet this constraint must posit that the non-cognitive attitudes that constitute moral judgements require a special social context in order to exist—they require, roughly, the existence of a network of entrusted concern relationships connecting all those who make moral judgements. That moral judgements presuppose such a network, and that such a network exists, are the central claims of fiduciary non-cognitivism.

The fiduciary non-cognitivist thus endorses a certain kind of externalism or anti-individualism about moral judgements. There is no ‘internal’ mental state, the fiduciary non-cognitivist claims, such that having that mental state necessarily amounts to holding a moral judgement.\(^45\) Inhabitation of an appropriate social setting—specifically, membership of an appropriate network of trust relationships—is an essential ingredient of moral judgement.

The idea that moral attitudes are dependent on social context is not new. Philippa Foot, as I described above, argued that attitudes of approval and disapproval can only

\(^{44}\) I discuss some possible counterexamples to this generalization involving descriptive judgement in 1.1.6. A weaker generalization excepting cases involving deception or inequalities of expertise would survive these counterexamples.

\(^{45}\) This is supposed to be true even if we hold the language of any candidate moral judgement-holder fixed. Thus the ‘externalism’ implied by fiduciary non-cognitivism is distinct from externalism about the contents of thoughts, of the sort that e.g. Tyler Burge’s famous arthritis example is supposed to establish (Tyler Burge, ‘Individualism and the mental’, *Midwest Studies in Philosophy* 4 (1979): 77).
exist in a certain kind of social context.\footnote{In her essay ‘Approval and Disapproval’, in Philippa Foot, \textit{Virtues and Vices and Other Essays in Moral Philosophy} (Oxford: Clarendon Press, 2002), 189-208.} Simon Blackburn, commenting on Foot, concedes that the same is true of the non-cognitive attitudes that constitute moral judgements.\footnote{Simon Blackburn, ‘The Flight to Reality’, in \textit{Virtues and Reasons: Philippa Foot and Moral Theory}, eds. Rosalind Hursthouse, Gavin Lawrence, and Warren Quinn (Oxford: Clarendon Press, 1995), 39.} However, we find no attempt in Blackburn’s writings to say what form such social contexts would take, or to determine, to any significant extent, the nature of their influence on our moral practices and psychologies. The novelty of the non-cognitivist theory defended in this dissertation lies chiefly in the centrality it gives to social institutions. According to fiduciary non-cognitivism, the network of trust relationships that makes moral judgement possible is fundamental to the explanation of the nature of moral judgement, of the nature and point of moral reasoning, of the practical significance of moral claims, and of the very existence of our moral practices. It is my contention that studying the social conditions of moral judgement, on a non-cognitivist construal, reveals the means to a unified explanation of many diverse features of our moral thought and practices.

I’ll now briefly say what fiduciary non-cognitivism is, and flesh out somewhat the version of the theory that I will be defending. Fiduciary non-cognitivism, as I have said, posits that moralists are parties to a network of entrusted concern relationships. These trust relationships oblige them to be responsive to one another’s preferences insofar as those preferences meet certain criteria. Preferences meeting these criteria are the non-cognitive attitudes involved in moral judgements. Each moralist is obliged to accommodate each other moralist’s relevant preferences unless her own relevant preferences are preferences for conformity to norms that prescribe doing otherwise. Fiduciary non-cognitivism about morality is, at its broadest, the species of non-cognitivism which holds that the non-cognitive attitudes involved in moral judgement are preferences which have this sort of significance—which oblige accommodation, or licence resistance to accommodation of others’ preferences—in the context of a network of entrusted concern relationships linking every moralist to every other moralist. I shall sometimes use the term ‘fiduciary non-cognitivism’ in a narrow sense, referring to the particular version of it that I defend in this thesis.\footnote{Why am I defending fiduciary non-cognitivism rather than fiduciary subjectivism? Why not posit that moral judgements are beliefs about people’s (idealized) preferences, where the...}
According to this particular version of the theory, the trust relationships that comprise the moral trust network all have the same terms. The preferences that are the non-cognitive components of moral judgements are preferences for conformity to norms. They concern which norms the moral judgement-holder herself should follow, and which norms other moralists should follow. The criteria that these preferences must meet guarantee that a moralist will have preferences, of the relevant sort, that she and others should follow exactly the same norms. Among the criteria imposed on the preferences that (partly) constitute moral judgements is a constraint on the desires that those preferences may be based on. With one exception, discussed below, the desires must all be broadly altruistic.

0.2.3  *Fiduciary Non-cognitivism and the Nine Desiderata*

I will now give highly abbreviated explanation of why my theory meets the ten desiderata, and in the process expand upon the brief summary of it that I have just given.

*Disagreement.* Fiduciary non-cognitivism was developed from the outset with the aim of explaining moral disagreement in mind. And, although its explanation of moral disagreement is not complete, I believe that fiduciary non-cognitivism meets this desideratum better than any rival non-cognitivist theory. It does so because it predicts that interpersonal clashes of moral judgement will elicit a stable rather than a fragile and impressionistic intuition of disagreement, at least where the judgements concern matters of basic moral principle. My theory of non-cognitive disagreement (developed in Chapters 1 and 2) predicts that such an intuition will be elicited by clashes of preference among parties to an entrusted concern relationship, and fiduciary non-cognitivism holds that clashing moral judgements involve clashing preferences of this sort.

*Justification, criticism, and demands for grounds.* Fiduciary non-cognitivism is well placed to answer Scanlon’s objection that communicating that one performed an action at the prompting of a certain non-cognitive attitude in no way answers an
objection to one’s performance of the action. According to the theory, when a moralist says that φ-ing is right (or that not-φ-ing is wrong), she communicates that she has a preference that she (and everyone else) should follow norms requiring φ. And the preference is purportedly one such that her possession of it establishes that (i) she can follow norms requiring φ without violating the trust of any other moralist who prefers that she should follow norms forbidding φ; and (ii) any other moralist who not-φs will violate her trust, unless he has a relevant preference to follow norms forbidding φ.

Thus when I am about to φ, and someone says to me, ‘φ-ing is wrong’, she communicates that she has a preference that I follow norms forbidding φ. Why should this preference be of any interest to me? Because, assuming it meets the relevant criteria, I have a trust-based obligation to acquiesce to it, unless I have a relevant preference to do otherwise. Because the speaker’s trust is at stake, the moral claim has the character of a demand, an objection, and a challenge to me to explain myself. When I respond by saying that φ-ing is the right thing to do, I communicate a preference that I should follow norms prescribing φ; if I genuinely have the preference, and it meets the criteria, this establishes that I did not violate the complainer’s trust by φ-ing. This is why my preference should be of interest to her. My moral claim has a justificatory character because it roughly amounts to a defence against an objection that I would violate her trust if I φ-ed.

The relevance of our preferences to the question of whether I have violated trust explains why my interlocutor and I should each have some sort of entitlement to demand to know the grounds for the other’s preference. I might want to know whether I really am obliged to acquiesce to my interlocutor’s preference absent a relevant preference of my own to do otherwise, perhaps because I am afraid that my own preference will not survive scrutiny. My interlocutor might want to know whether my expressed preference really exists and gives me discretion to resist her demands. Explaining why we hold our judgements is a way of establishing whether the preferences they involve meet the relevant criteria, and so of establishing whether I will violate trust if I φ. Since the information is of significant interest to us, it makes sense that the terms of our trust relationship should in some way entitle us to demand of each
other that we should explain our moral judgements when we use them to criticize or justify.\textsuperscript{49}

\textit{Moral uncertainty}. Fiduciary non-cognitivism gives an account of moral uncertainty that makes plausible predictions about the relationship between moral uncertainty and action.\textsuperscript{50} More specifically, it gives an account of degrees of confidence in moral claims, and an account of moral indecision. The account is impossible to summarize here without considerable distortion. But very roughly, my \textit{confidence} in a moral claim is the expected utility of conformity to it—that is, the expected utility in terms of the satisfaction of my altruistic desires (and my ‘simplicity desire’, discussed below). If I am morally \textit{undecided} about a moral claim \textit{J}, the expected utility of conformity to it, in comparison to the expected utility of conformity to its negation \textit{~J}, is \textit{indeterminate}—perhaps because my relevant credences are indeterminate or imprecise. The result of this indecision is that I lack a preference between conforming to \textit{J} or \textit{~J}. This in turn means that I cannot act on \textit{J} or \textit{~J} without violating the trust of other moralists.

\textit{Moral reasoning and changes of motivation}. As I noted above, explaining how moral reasoning can produce changes in motivation is a problem for moral psychologies that postulate that the motivations affected are \textit{basic} or non-instrumental. Fiduciary non-cognitivism is unusual among non-cognitivist theories in that it does not make this postulation. The theory characterizes reasoning about basic moral principles as a kind of means-end reasoning: moral reasoning is a quest to identify the set of norms conformity to which would best satisfy the set of desires specified by the terms of the moral trust relationships—chiefly altruistic desires, but also a desire that moralists should follow simpler rather than more complex norms. The process involves descriptive reasoning, which, for rather complicated reasons, is almost exclusively \textit{a priori}. This descriptive reasoning concerns the consequences of following certain norms in certain possible circumstances (e.g. whether there are possible circumstances in which following utilitarian norms would mean committing judicial murder).

\textsuperscript{49} Or perhaps the terms of the trust relationship don’t need to stipulate this entitlement; perhaps, as a general matter, if someone makes a trust-based claim of me, she has a trust-based obligation to be cooperative if I am trying to establish whether she really has the claimed entitlement. If so, this might explain my entitlement to demand that she supply grounds for her judgement.

\textsuperscript{50} It does not offer a corresponding semantics, however, to explain e.g. the significance and logical properties of sentences like ‘Abortion is probably wrong’.
What explains the reliable changes in motivation that result from changes in basic moral judgement is a moralist’s desire to avoid violating other moralists’ trust. If a moralist ever acts in violation of her preferred norms, she violates the trust of any moralist who prefers that she should do otherwise. Since there is sure to be some moralist whose trust she would violate in this way if she acted in violation of her preferred norms, her desire not to violate trust will bid her conform to these norms, whatever they may prescribe. A person must desire to live up to the trust of all other moralists if she is to be a member of the moral trust network, and she must be a member of the moral trust network if she is to be a moralist herself.

Why morality exists. Fiduciary non-cognitivism makes what on the face of it is a wild empirical postulation, namely, that everyone in the universe who makes moral judgements is party to a rather complicated sort of trust relationship with everyone else in the universe who makes moral judgements, and that all of these trust relationships have nearly exactly the same terms. But it is at least possible to tell an evolutionary story that will explain why humans are disposed to form these sorts of trust relationships. The evolutionary story I offer is not independently plausible, and it is based purely on inexpert conjecture. But for all its scientific irresponsibility, it posits nothing as immodest as the more-or-less ad hoc hypotheses, made by all other psychologically detailed non-cognitivist theories, concerning our evolved propensities to form certain mental states when we have certain other combinations of mental states; or Gibbard’s hypothesis in Wise Choices, Apt Feelings that we evolved a novel, distinctively normative and ‘linguistically infused’ kind of mental state, viz. norm-acceptance. My story requires only much more superficial changes in the way our brains are wired.

The postulation of a disposition to form moral trust relationships is made less costly by the plausibility of the proposition that we do form entrusted concern relationships—the sorts of trust relationships connecting all moralists according to fiduciary non-cognitivism—in a tacit and spontaneous way in our interactions with other people outside of normative contexts. I argue for this proposition in the course of defending my theory of non-cognitive disagreement. I also explain why it would, in some non-moral contexts, serve people’s interests to form networks of entrusted concern relationships, like the moral trust network but confined to small interacting groups with shared ends. These networks would, I argue, share properties with the moral trust network that we might otherwise have thought were peculiar to the latter.
These arguments, I hope, convincingly assimilate moral thought and talk on a fiduciary non-cognitivist construal to thought and talk of a more humdrum and familiar kind, a kind involved in a many mundane instances of practical negotiation and joint deliberation. Since we are manifestly capable of engaging in these ordinary kinds of negotiations and deliberations, and since a construal of these negotiations and deliberations as presupposing tacit entrusted concern relationships between the parties is plausible, we should not find the costs of positing the evolution of dispositions to form the moral trust network altogether overwhelming. We evidently possess a large part of the cognitive machinery we would need to have if this postulation were true.

There are devils in the details of my theory. Even if we are prepared to countenance a universe-spanning moral trust network, fiduciary non-cognitivism still has some empirically troublesome surprises in store on close scrutiny. For one thing, it requires moralists to effortlessly form preferences and make decisions based on fantastically small credences (3.3.3). It is not antecedently clear that we have the cognitive resources to do this.51 For another, it requires moralists to have certain desires which we do not obviously have, namely, altruistic desires and ‘simplicity desires’—which, again, are desires that moralists should follow simple norms. Evolutionary mechanisms such as kin selection, reciprocal altruism, and group selection could credibly account for our possession of the posited altruistic desires. But I have, as of writing, failed to come up with an explanation for our possession of simplicity desires. These are certainly serious shortcomings of my theory, as it is presented in this work, but they are only evident because my theory of our moral thought and practices, and my explanation for their existence, is worked out more fully and explicitly than is typical for non-cognitivist theories. There are devils in the details of my accounts of moral disagreement and substantive moral reasoning because the accounts are detailed—more so than the rival accounts of Gibbard, Blackburn, and Ridge. It would be a mistake to reject fiduciary non-cognitivism because of its empirically costly postulation that moral reasoning involves fantastically small credences, and its empirically mysterious postulation of simplicity desires, before it has been shown that rival theories fare better when they have joined their dots to the same extent.

51 The rational outcomes of this instrumental reasoning involving minuscule probabilities could be arrived at by easy shortcuts, but if I posit that we typically exploit these shortcuts, I face pressure to posit that we evolved a special disposition to do so, which is also costly.
Ubiquity of moral thought and talk. As I have already more or less acknowledged, the fact that people almost everywhere seem to be capable of making moral judgements, whether or not they have interacted in the past, is a special problem for fiduciary non-cognitivism. This is, of course, because fiduciary non-cognitivism posits that one cannot make moral judgements unless one shares trust relationships with everyone else who can make moral judgements, and it is prima facie implausible to suppose that geographically separated and non-interacting pairs of moralists will have formed the requisite trust relationships. A large portion of this thesis will be devoted to showing that meeting this desideratum is not so great a problem for fiduciary non-cognitivism as it appears.

Rule-like character of moral requirements. Explaining our tendency to think of moral requirements as consisting of general rules is a challenge for fiduciary non-cognitivism. The thoughts that put parties to non-moral entrusted concern relationships in disagreement are usually just ad hoc preferences concerning what someone is to do in a particular situation, and are not obviously voiced with a presumption that they will guide the agent’s decisions in similar circumstances in the future. So fiduciary non-cognitivism must explain why its posited moral trust relationships are different.

I explain the rule-based character of morality as arising from moralists’ need to be able to determine, as easily as possible, whether other moralists are abusing their trust by misrepresenting their true preferences. Such abuses can sometimes be detected by comparing the moral claims a person has made on different occasions, provided the actions in question are sufficiently similar. If there is a constraint on how complex the totality of a person’s moral judgements may be, a disparity between two claims made on different occasions will be stronger evidence of abuse, even where the actions in question are somewhat different. Thus it makes sense that there should be some sort of simplifying influence or constraint upon moral judgements. The most straightforward way to realize this, given that fiduciary non-cognitivist’s account of the nature and function of moral judgements is broadly true, is as follows: those preferences of moralists that oblige accommodation and licence discretion should be preferences concerning which rules to follow, and they should be based in part upon desires that moralists should follow simpler rather than more complex rules.

Moral forgetting. I argue that the best explanation for our intuitions about moral forgetting is provided by certain types of non-cognitivism, for the reason given in 0.2.1,
and that fiduciary non-cognitivism fares no worse on this score than other non-cognitivist theories.

*Moral approval and disapproval.* Fiduciary non-cognitivism allows us to give an account of moral approval and disapproval that is consistent with Foot’s hypothesis that approval and disapproval depend on social context. The fiduciary non-cognitivist holds that the social context needed for approval and disapproval is an entrusted concern relationship between the approver or disapprover and the person of whose actions she approves or disapproves. Arguably, anyone is capable of moral approval and disapproval, and anyone who, or whose token actions, can be morally approved or disapproved of, will be someone who is capable of making moral judgements. Because fiduciary non-cognitivism holds that membership of the moral trust network is a prerequisite for moral judgement, the fiduciary non-cognitivist predicts that anyone who morally approves or disapproves will be a party to an entrusted concern relationship with the approved or disapproved party. So she can hypothesize that such a relationship is what explains a person’s capacity to morally approve or disapprove, in all cases of moral approval and disapproval.

### 0.2.4 Scope and Omissions

The theory of moral reasoning and moral psychology that I develop in this thesis is incomplete. For the purposes of evaluating my theory, the questions and issues that I might have addressed but haven’t fall into three categories: (1) those that I have not tried to address because my ambitions are limited, and which do not obviously pose a special problem for fiduciary non-cognitivism; (2) those that I have not tried to address, and which do obviously pose a special problem for fiduciary non-cognitivism; and (3) those that I have tried but failed to address.

The most important issue falling under (1) concerns moral language and deductive reasoning. Non-cognitivists have been excoriated for devoting their attention to grand metaphilosophical and programmatic questions, and traditional big-picture metaethical

---

52 Admittedly, it is not clear that we cannot morally disapprove of the actions of a professed amoralist. The fiduciary non-cognitivist may be forced to make the hard-to-test postulation that this is because we cannot shake the idea that deep down she is still a moralist. In support of this, it does at least require an unusual effort to imagine a normally socialized human who would not sincerely complain or at least feel wronged or appalled if she and her family were being subjected to some extreme cruelty, however adamantly she may have professed to amoralism when there was less at stake. It also seems plausible that one cannot morally disapprove of the token actions of an animal or an unsocialized ‘feral’ human.
questions, while neglecting to produce a detailed and empirically adequate theory of moral language.\textsuperscript{53} It may thus seem perverse that I have chosen to ignore linguistic questions entirely in this dissertation, and focus on broad questions about our moral psychology and practices. Although I developed my non-cognitivist theory with a certain strategy for solving the Frege-Geach problem in mind, carrying out that strategy turned out to be more than I could attempt in the present work. Predicting that moral judgements are capable of disagreeing, though doubtless easier, is as vital to the adequacy of a non-cognitivist theory as predicting that people use moral words in the way that they do. And when I turned my attention to the issue of disagreement, I found that a non-cognitivist could satisfactorily account for the phenomenon only by postulating a ‘moral trust network’, or something similarly extraordinary. Determining the costs and significance of such a postulation turned out to be the work of an entire dissertation. Doing this work seemed a more worthwhile contribution to non-cognitivist theory than attempting to solve the Frege-Geach problem, because it seemed less likely that other theorists would turn their attention to it.

My linguistic reticence has required some psychological imprecision. If I were to give an account of moral language, it would be a broadly expressivist one. Since, loosely speaking, expressivists hold that moral sentences serve to express the non-cognitive attitudes that constitute moral judgements, and that the logical properties of sentences are explained by the rational properties of moral judgements, my silence about moral language requires me to be silent about the exact psychological nature of moral judgements. So—disappointingly perhaps, given that moral psychology is my primary focus—I have not given a complete account even of simple judgements that actions are permissible or impermissible, let alone conditional and other complex moral judgements. For similar reasons, though I have said quite a bit about what substantive moral reasoning involves, I have omitted to give an account of deductive moral reasoning.\textsuperscript{54}

My failure to provide an account of moral language compatible with non-cognitivism falls under (1) because, of course, the need to provide such an account is not peculiar to fiduciary non-cognitivism. Here are some other omissions in the same

---

\textsuperscript{53} See the Preface to Schroeder’s \textit{Being For}.

\textsuperscript{54} I do, however, offer an explanation for the irrationality of having basic moral principles that both permit and forbid an action. I could not have said much about moral disagreement without addressing this topic, and explaining moral disagreement is one of the central concerns of this work.
category. My account of moral judgement—incomplete as it is—applies only to narrowly normative judgements concerning actions; I say nothing at all about judgements of good and bad, and judgements concerning the propriety of emotions. I have not attempted to explain judgements concerning the gravity or weight of moral requirements, nor to account for judgements of pro tanto wrongness and requirement. I provide no account of blame and other reactive attitudes. My explanation of moral disagreement only concerns disagreement about what to do in possible future choice situations, not disagreement about past decisions or purely hypothetical decisions. I make no attempt to explain how the broad account of moral psychology and practices I defend could be extended so as to explain other kinds of normative thought and practices.

Let us turn to category (2), which includes questions that I need to address if I am to show that fiduciary non-cognitivism does not face special problems, and which I have neglected to address in this dissertation. First, I do not address the question of whether we need to appeal to moral-psychological facts to explain the phenomenon of trust. The viability of fiduciary non-cognitivism, at least in the form defended here, requires a negative answer to this question. Second, I do not fully address questions concerning

---

55 Margaret Walker makes the following observation which will be familiar to those acquainted with the literature on trust: if I expect you to do something that benefits me because you have often done so in the past, and not because you have given any assurance that you will do so, I do not (if I am reasonable) trust you to continue doing it. If you stop, I will not feel resentful or respond reprovingly, as I might if my trust had been disappointed. Walker diagnoses this as follows: ‘I’m only prone to these reactive feelings that blame and punish … when I rely on you to do what you should. I do not only expect that you will do it, I expect it of you. My expectation is a normative expectation. Normative expectations are not simply confident assumptions that people are likely to behave in particular ways; these expectations express instead a stance toward others that demands certain behavior of them, because it is what they are supposed to do.’ Margaret Urban Walker, Moral Repair: Reconstructing Moral Relations After Wrongdoing (Cambridge: Cambridge University Press, 2006), 79-80. Setting aside Walker’s claim about ‘demanding stances’, this seems quite plausible to me, but as yet I see no reason to think that, if Walker is right, the normative judgements that are necessarily involved in trust are moral judgements, or normative judgements of any kind that is similarly perplexing to philosophers. They may just be judgements about what is required by the social norms that you and I are ‘in the grip of’, to invoke an idea of Gibbard’s (Wise Choices, 60). Assurances, promises, and the like, all create obligations according to social norms. If we take judgements of social-normative obligation to explain both a person’s ‘demanding stance’ toward another and (hence?) the possibility of trust between the two, we do not commit to anything insurmountably problematic for fiduciary non-cognitivism. (This view would commit the fiduciary non-cognitivists to the idea that humans are ‘hard-wired’ to be in the grip of the terms of the moral trust network, and to expect all other humans to be in their grip, or something along these lines.)

56 If it turned out that trust does depend on normative judgement of a problematic kind, it may be possible to provide a (quasi-)fiduciary non-cognitive analysis of moral judgment employing
the possibility and implications of failures of trust among moralists, and of deliberate withdrawal from the moral trust network. I discuss some such questions in 5.4, but I have failed to address others—e.g., why doesn’t someone cease to be a proper object of moral evaluation if she publicly declares her defection from the moral trust network?

As for category (3) there is, to my knowledge, one main respect in which my theorizing has fallen short of my aims. As I acknowledged in 0.2.3, I have failed to provide an explanation of moralists’ simplicity desires—of why moralists desire conformity to simpler norms. Thus my attempt to show that a non-cognitivist can explain what is going on when we engage in substantive reasoning about basic moral principles without positing an important role for higher-order desires is not a complete success.

0.2.5 Method

I take my main project in this thesis to be one of conceptual analysis. I present the beginnings of an analysis of the concept of moral judgement (or rather, the concepts of judging that φ-ing is morally wrong and judging that φ-ing is morally permissible) and some related concepts, which aspires to explain our intuitions about moral thought and talk. However, a very large portion of this work is devoted to defending an empirical hypothesis, namely, that the moral trust network I describe exists. My defence of this claim is part of my defence of the conceptual analyses. Philosophers are perhaps not accustomed to thinking of the plausibility of conceptual analyses as depending on empirical propositions, aside from facts about language use and folk intuitions. So my preoccupation with this empirical question may require some explanation.

The explanation is simple. If the things that fall under a concept play, or seem to play, an important role in our lives, an analysis of that concept will be less plausible if a credible explanation cannot be given as to why things that fall under the analysans play, or seem to play, an important role in our lives, even if it is only a contingent fact that they do so. Moral judgements seem to play an important role in our lives: we seemingly try to conform our behaviour to our moral judgements, we seemingly take pains to


57 To avoid disappointment, I should say now that this defence will not draw on scientific evidence, but merely on common sense judgements about hypotheses and imaginary cases. I will say more about this approach and its reliability in the introduction to Chapter 3.
reason well when arriving at our moral judgements, we seemingly try to bring other people’s moral judgements into alignment with our own, sometimes through impassioned argument, and so on. There must be a reason for their seeming importance, and an analysis of moral judgements has implications for what that reason could be. For instance, if the analysis states that moral judgements are beliefs about what the Judaeo-Christian God forbids, it will be hard to explain why moral judgements seem to have played an important role in the lives of people who have never heard of the Judaeo-Christian God. So my sketchy analysis of moral judgements as (in part) voiceable preferences of parties to a moral trust network will be less plausible if I can’t show that it is compatible with a plausible explanation for the appearance that such preferences matter to us.

One way to try to explain this would be to argue that we actually have these voiceable preferences, that we are tacitly aware that we have them, and that they would matter to us if we were aware that we had them. This approach would, if successful, vindicate the appearance that moral judgements play an important role in our lives. Another approach would be to allow that moral judgements perhaps don’t exist, and defend an error theory as to why they seem to play an important role in our lives nonetheless. I’ve chosen the vindicatory approach, because it is straightforward, and because I can’t think of an error-theoretic approach that would be more plausible. Both of these approaches seem to require defending empirical claims, because they are approaches to explaining a contingent phenomenon—the appearance that moral judgements seem to play an important role in our lives.

0.2.6 A Look Ahead

I’ll now summarize the contents of this dissertation. The work has two parts. Part I, ‘Trust and Non-Cognitive Disagreement’, consists of Chapters 1-3, and concerns non-cognitive disagreement in general and the relationships of trust that make it possible. In Chapters 1 and 2 I defend a theory of non-cognitive disagreement. Chapter 1 concerns disagreement between two people about what one of them is to do. Chapter 2 concerns disagreement between two people about what a third should do. In Chapter 3 I consider networks of entrusted concern relationships, why they would exist, and what forms they would take under normal circumstances. I argue that several of the apparently distinctive features of the moral trust network that fiduciary non-cognitivism posits are ones that we should expect to find in any trust network formed under circumstances that
are in certain respects similar, and I provide what I hope is a natural-seeming example of such a network.

Part II, ‘Moral Psychology and the Moral Trust Network’ consists of Chapters 4-9. It is focused specifically on describing and defending my theory of moral thought and practices. In Chapter 4 I present fiduciary non-cognitivism’s account of the non-cognitive attitudes that are involved in moral judgements of requirement, permissibility, and impermissibility, and present an explanation—somewhat limited in scope—of moral disagreement. Chapter 5 addresses my sixth metaethical desideratum (‘ubiquity of moral thought and talk’). I try to explain how a network as extensive as the moral trust network could exist, and give some reason for thinking that it does. Among other things, I argue, drawing on the observations of Foot that I discussed earlier in this introduction, that we must postulate something like the moral trust network if we are to explain the phenomena of moral approval and disapproval (i.e. to meet my ninth desideratum). In Chapters 6, 7, 8, and 9 I attempt to show that fiduciary non-cognitivism can meet my seventh, third, fourth, and eight desiderata (the ‘rule-like character of moral requirements’, ‘moral uncertainty’, ‘moral reasoning and changes of motivation’, and ‘moral forgetting’) respectively, and in some cases that fiduciary non-cognitivism does so better than rival theories.
PART I

TRUST AND NON-COGNITIVE DISAGREEMENT
1

Conditions for Robust Non-cognitive Disagreement

Non-cognitivists about morality hold, roughly, that what we call ‘moral beliefs’ are not beliefs on an orthodox philosophical understanding but non-cognitive or desire-like attitudes. Thus they hold that moral disagreements are non-cognitive disagreements, involving ‘clashing’ desire-like attitudes of some sort rather than opposing cognitive beliefs. Moral disagreement (and normative disagreement in general) has a property, which I call ‘robustness’, that not all ‘clashes’ of non-cognitive attitudes possess. If the non-cognitivist theory of moral psychology that I develop in this thesis is to explain the robustness of moral disagreement, it must identify moral disagreement with a kind of non-cognitive disagreement that is robust. In this chapter, I try to identify necessary and sufficient conditions for robust non-cognitive disagreement, in a certain range of cases. The theory of robust non-cognitive disagreement that I develop in this chapter applies only to cases in which two parties disagree about what one of them is to do, or about what both of them are to do. I extend the theory to accommodate cases in which two people disagree about the choices of a third party in Chapter 2. My theory draws heavily on Philippa Foot’s observations concerning the social conditions for approval and disapproval.58

In Section 1.1 I describe the phenomenon of ‘robust’ non-cognitive disagreement that I hope to explain; give examples of it; and evaluate some candidate explanations for the phenomenon, or accounts of the conditions for its occurrence, which I find wanting. In Section 1.2 I present and defend my favoured explanation.

58 I say more about my theory’s debt to Foot in 5.3.1.
1.1 The Phenomenon and Some Explanations

1.1.1 Robust Non-cognitive Disagreement

It is easy to think of cases in which different people seem to disagree, in some sense or other, in virtue of their non-cognitive attitudes. Unfortunately, of all such cases, few (if any) elicit intuitions of disagreement that are as strong and resilient as those elicited by straightforward cases of normative disagreement. We may be inclined to think, with Brian Weatherson, that ‘If A says “I like ice cream” and B says “I don’t like ice cream”, then there is a natural sense in which they are disagreeing’. But this is surely a contestable matter; it would not be unreasonable to think that A and B were not disagreeing.

We can bring this out by imagining that A and B’s exchange of liking-reports is followed by a discussion of whether or not they disagree:

**Liking Conciliation**

A: I like ice cream.
B: I don’t like ice cream.
A: We disagree then.
B: I don’t think we disagree. We just like different things.

B makes what I shall call a conciliatory claim—a claim that their difference with respect to ice cream is something other than a disagreement. The conciliatory claim seems natural and appropriate in this context. By contrast, conciliatory construals of opposing normative (and evaluative) judgements are inappropriate:

**Nice Conciliation**

A: I believe that ice cream is nice.
B: I believe that ice cream is not nice.
A: We disagree then.
B: I don’t think we disagree. We just believe different things.

B’s conciliatory claim seems simply false. There is no plausible way to construe A and B’s reported difference of attitudes toward ice cream in this example as anything other than a disagreement (assuming their reports are sincere). Among those differences of attitude that (as they are described or reported) might be construed as disagreements, I shall call those whose status as disagreements cannot plausibly be denied—i.e. those that make conciliatory claims inappropriate—robust disagreements.

---

Ordinary normative disagreements seem to be robust, and a non-cognitivist account of normative disagreement ought to predict this. So the non-cognitivist must model normative disagreement on a species of non-cognitive disagreement that is not susceptible to conciliation. I think some of Stevenson’s famous cases of ‘disagreement in attitude’ provide such a model:

Suppose that two people have decided to go to a restaurant together. One suggests a restaurant where there is music; another expresses his disinclination to hear music and suggests some other restaurant. It may then happen, as we commonly put it, that they ‘cannot easily agree on which restaurant to choose.’ … Mrs. A has social aspirations, and wants to move with the elite. Mr. A is easy-going, and loyal to his old friends. They accordingly disagree about what guests they will invite to their party.  

Let’s test the robustness of these disagreements. Imagine the following exchange between the prospective diners.

Restaurant

A: Let’s have dinner together.
B: Okay. Where do you suggest?
A: I want to go to Minims. There’s a good piano trio playing there tonight.
B: Do you? Because I’m not in the mood for music. I’d prefer somewhere quiet.
A: We disagree then.
B: I don’t think we disagree. We just prefer different things.

Conciliation seems inappropriate here. Intuitively, it is made inappropriate by the fact that A and B have a decision to make, a decision to which each party’s preferences are clearly relevant. There is a matter that must be resolved between them, and it remains stubbornly unresolved after B’s attempt at conciliation.

Now imagine the following exchange between Stevenson’s Mr. and Mrs. A:

Invitation

Mr. A: I’d rather we invited Mr. Chum. He’s an old friend.
Mrs. A: But as you know, Mr. Toff’s connections are excellent. I want to invite Toff.
Mr. A: We disagree then.
Mrs. A: I don’t think we disagree. We just prefer different things.

Again, conciliation seems inappropriate. Intuitively, Mr. and Mrs. A are in disagreement over whom to invite, and pointing out that this is owing to a mere difference in

60 Charles L. Stevenson, Ethics and Language (New Haven; Yale University Press, 1944), 3. Stevenson describes two further disagreements, one between a museum curator and his advisor, and another between a mother and her son. I don’t discuss the museum case I have found it difficult to present in such a way that we reliably interpret it as a disagreement in preference rather than belief. I discuss the mother and son case below.
preferences does not reveal it to be any less of a disagreement. Their difference in preferences must be a source of contention as long as it remains to be decided whom they will invite to the party. Their disagreement is robust.

1.1.2 Type-of-Attitude Explanations

Why are the non-cognitive disagreements Stevenson describes robust? It might be thought that this is explained by the nature of the opposing non-cognitive attitudes in Stevenson’s cases. The parties in Stevenson’s cases seem to have conflicting preferences, in contrast to the clashing likes and dislikes reported in Weatherson’s ice cream case. Perhaps preferences held by different people disagree whenever they are mutually unsatisfiable, or could not consistently be held by the same person. But the following example from Michael Ridge shows that mutually unsatisfiable preferences do not suffice for robust disagreement.

Flat (Preferences). A flat $F$ is available to rent, and both $A$ and $B$ want to rent $F$. (That is, there is a time $t$ such that $A$ prefers that $A$ rather than $B$ should rent $F$ at $t$, and $B$ prefers that $B$ rather than $A$ should rent $F$ at $t$.)

A and B have preferences that are mutually unsatisfiable, yet they do not seem to disagree. Here is a different sort of counterexample, involving a conciliatory claim.

Retail

*Shopkeeper:* Do you want to buy this coat, sir?
*Customer:* No. I’d prefer something cheaper.
*Shopkeeper:* Well, I’d rather you bought it. We disagree.
*Customer:* Pardon?
*Shopkeeper:* I want you to buy the coat, you prefer not to. We disagree.
*Customer:* I don’t think we disagree. We just prefer different things.

Customer’s conciliatory claim seems appropriate here. An interpersonal clash of preferences does not suffice for robust disagreement.

Another possibility is that robust non-cognitive disagreement is disagreement in plans or intentions. This idea derives some appeal from the observation that plans are more closely bound to practical deliberation than preferences are. You can have preferences for states of affairs (e.g. that Harold had won the Battle of Hastings) the realization of which cannot be the aim of your practical deliberations; the same is not true of plans. Stevenson’s cases—at least those I have quoted—all plausibly involve

---

61 This is borrowed (not quoted) from Michael Ridge, ‘Disagreement’, *Philosophy and Phenomenological Research* 86, no. 1 (2013): 46.
practical decision-making in one way or another, and it is prima facie plausible that the contending parties all disagree because they have mutually unrealizable plans. Nonetheless, it is a straightforward matter to modify Ridge’s flat example so that it will serve as a counterexample to this proposal:

*Flat (Plans).* A and B both intend to rent the same flat at the same time. (That is, there is a time \( t \) such that A intends that she rather than B should rent it at \( t \), and B intends that he rather than A should rent it at \( t \).)

A and B have plans that are mutually unrealizable, yet do not seem to disagree. I think that we can also imagine a version of *Retail* in which Shopkeeper reports an intention that Customer should buy the expensive coat, Customer reports a contrary intention, and Customer makes a felicitous conciliatory claim.

There is a class of attitudes that seem to be more complex than attitudes like desiring, intending, or liking, and which seem to constitute excellent materials for a type-of-attitude explanation of the phenomenon of robust non-cognitive disagreement. This class includes attitudes such as approving and disapproving, being for and being against, favouring, endorsing, and opposing. The parties in Stevenson’s cases certainly seem to favour, and to be for, incompatible options. If I approve and you disapprove of the ‘Oxford comma’ in scholarly prose, or ‘emoticons’ in emails, who would deny that we disagree? And it is somehow odd or superfluous to think of either party in Ridge’s flat case as being against or opposed to the other party’s getting the flat, if they are merely in competition for it.

However, I have reservations about relying on these materials for my theory of robust non-cognitive disagreement, because attitudes of approval etc. are not well-understood mental states. There are widely-held and well-scrutinized views about the nature of desires and preferences—the dispositional theory of desires for instance, and the thesis that, necessarily, \( S \) prefers \( p \) to \( q \) just in case \( S \) desires \( p \) more than \( q \) (whichever side of this biconditional one takes to be explanatorily fundamental). Intentions are more mysterious, but they have at least received a good deal of philosophical attention, and we have some idea of the available theoretical options and

their merits. But approval, favouring, etc., are undertheorized, and the fact that they can put their possessors in robust disagreement itself demands an explanation. Even if it gets the facts right, an explanation of robust disagreement as the result of clashing attitudes of this sort, absent an account of those attitudes themselves, would be superficial and a case of *obscurum per obscurius*.

Since my aim is ultimately to explain moral disagreement, it is especially worrisome that I cannot rule out the possibility that approval etc. involve or are explained by normative judgements. Suppose that I hypothesized that robust non-cognitive disagreements consisted in clashing attitudes of approval and disapproval. And suppose that it turned out that to approve of \( \varphi \)-ing fundamentally involved judging that \( \varphi \)-ing was right, and that to disapprove of \( \varphi \)-ing involved judging that \( \varphi \)-ing was wrong. My hypothesis could not then, in conjunction with the thesis that moral judgements are attitudes of approval and disapproval, explain our intuition that two people disagree when one judges that \( \varphi \)-ing is right and the other that \( \varphi \)-ing is wrong.

### 1.1.3 Attitudes, Plus Attitudes about Attitudes

Writers on non-cognitive disagreement often find that something more is needed to capture our intuitions of disagreement than the possession by different people of attitudes that are mutually inconsistent (or attitudes that a single person could not rationally hold simultaneously). A good theory of non-cognitive disagreement will, it is felt, predict that the disagreeing parties are in some sense in conflict. Disagreeing parties, one might suppose, should not be utterly indifferent to each other’s attitudes; neither should be entirely happy for the disagreement to persist; and each should prefer

---

64 I do present the beginnings of an analysis of approval and favouring in 3.3.2.
66 If I provided an analysis of one of these complex attitudes (and the analysis did not use ‘genuinely’ normative terms), then relying on it in my theory of robust non-cognitive disagreement would be unproblematic. Arguably this is what I end up doing. According to my theory, robust non-cognitive disagreements are clashes of voiceable preferences, and voiceable preferences (i.e. preferences that we *voice* in decisions, as opposed to merely reporting) might be thought of as ‘complex’ attitudes like approval. Much of the work of defending this theory can be thought of as a defence of my analysis of a voiceable preference, since it is already intuitively plausible that if two people voice opposing preferences in a decision, they disagree. (The incomplete analysis of approval that I present in 3.3.2 does not provide a shortcut to an explanation of robust non-cognitive disagreement, since it depends on my account of voiceable preferences.)
that the disagreement end with a ‘concession’ on the other’s part, rather than on her own.

These ideas motivate the following theory of robust non-cognitive disagreement.

**Mock-Stevensonian Hypothesis.** A and B are in robust non-cognitive disagreement iff there is some agent S and some φ such that (i) A prefers that S should φ, (ii) B prefers that S should not-φ, (iii) A prefers that B should not prefer that S not-φ, and (iv) B prefers that A should not prefer that S φ. (S may be a group agent, and may be the same agent as A and/or B.)

This seems to make the right predictions in Stevenson’s cases. I call it the Mock-Stevensonian Hypothesis because it is inspired by Stevenson’s conditions for ‘disagreement in attitude’:

> Two men will be said to disagree in attitude when they have opposed attitudes to the same object—one approving of it, for instance, the other disapproving of it—and when at least one of them has a motive for altering or calling into question the attitude of the other.68

The Mock-Stevensonian Hypothesis rather than Stevenson’s conditions will be the target of my counterexample, because I’m not sure what Stevenson means by ‘calling into question’ attitudes, and because Stevenson does not purport to be giving conditions for what I am calling robust non-cognitive disagreement. Here is the case:

**Rival Hostesses.** Mrs. A prefers that Mr. Toff should attend her party at t; Mrs. B prefers that Mr. Toff should attend her party at t; neither wants competition, so each prefers that the other should prefer otherwise than she does.

The Mock-Stevensonian Hypothesis predicts that Mrs. A and Mrs. B are in robust non-cognitive disagreement, but they do not seem to disagree. The hostesses’ interest in changing each other’s preferences does not suffice to turn their clashing preferences into a robust disagreement.

### 1.1.4 Is ‘Robust Non-Cognitive Disagreement’ Really Non-Cognitive?

I have already observed that normative disagreements are robust. Perhaps the parties in Stevenson’s cases seem to disagree robustly because we attribute disagreements in normative belief (understood as a cognitive attitude) to the parties. Call this the **Normative Disagreement Hypothesis.** On the most obvious version of this view, parties to an apparent non-cognitive disagreement disagree about what someone ought to do, and, if their disagreement takes verbal form, communicate their beliefs by reporting their preferences. They can communicate their ought-beliefs in this way because of our

---

expectation (at least for some subset of our ought-beliefs) that someone who believes
that an agent ought to φ will prefer that that agent φs.

How well does the Normative Disagreement Hypothesis explain our judgements
about the cases we have considered? It is quite plausible that Mr. and Mrs. A disagree
about whom they ought to invite to their party, and it is at least superficially plausible
that the parties in Restaurant disagree about where they ought to dine. These cases seem
to support the hypothesis. By contrast, we are not much inclined to think that the parties
in Flat (Preferences), Retail, and Rival Hostesses have normative beliefs of a sort that
would put them in disagreement.

It is not immediately clear why we attribute relevant normative beliefs to both
parties in Stevenson’s cases (assuming that we do), but not to the parties in these other
cases. This may be partly because we assume that people are more likely to go to the
trouble of forming normative beliefs if they want to persuade other people to do things.
This would explain most of the data: the parties in Stevenson’s cases seem interested in
persuading each other to do things, whereas those in Flat and Rival Hostesses do not.
(Retail is, I think, more complicated, but still raises no problems for the hypothesis. 69)

So far, so good. However, on closer inspection, the Normative Disagreement
Hypothesis does not seem to give a credible explanation for our judgement that the
parties in Restaurant robustly disagree. I take the following to illustrate the kind of
disagreement the hypothesis requires us to attribute to the parties in this case:

69 Shopkeeper seems interested in persuading Customer to buy the expensive coat. Why don’t
we attribute to Shopkeeper a belief that Customer ought to buy the coat, and to Customer a
belief that it is not the case that he ought to buy the coat? Our explanation will differ depending
on the flavour of the ‘ought’ in question. Consider the hypothesis that we attribute beliefs to the
parties about whether Shopkeeper has a claim upon Customer that he buy the coat—the ‘ought’
thus has the flavour of an interpersonal demand. On this view it is fairly plausible that we
attribute a relevant ought-belief to Customer, namely, a belief that Shopkeeper has no such claim
upon him. But we plausibly don’t attribute to Shopkeeper a belief that Customer ought to buy
the coat in this sense of ‘ought’ because this proposition is incredible. Now consider the
hypothesis that we attribute beliefs about a less interpersonally demanding sort of ‘ought’ to the
parties. Perhaps we attribute to Shopkeeper a belief that Customer ought to buy the coat because
it would be a fine thing if Customer possessed it and a pity if he did not (a belief of the sort she
might express by saying, ‘You simply must buy this coat!’); and perhaps we attribute to
Customer a contrary opinion. Now it is more plausible to suppose that we attribute to
Shopkeeper the relevant belief (although: (i) it is not really the sort of normative belief one
would communicate by reporting a preference, and (ii) we would have strong reason to doubt
his sincerity if he did). But it is less plausible to suppose that we attribute a relevant belief to
Customer. She might simply not have gone to the trouble of forming such an opinion, because
she had no interest in persuading Shopkeeper to do anything.
**Restaurant (Ought)**

A: We ought to go to a restaurant with music because I feel like listening to music.

B: I disagree. We ought to go to a restaurant without music because I don't feel like listening to music.

This interpretation seems strange because the grounds the parties offer for their ought judgements are self-regarding. It is strange to suppose that either party is prescribing that they both should go to her preferred restaurant, when she has nothing better to say in favour of her prescription than that it is the restaurant she feels like going to, even though she knows that her interlocutor has an opposing inclination. This seems selfish and arrogant. And it is not plausible that our intuition that Stevenson’s parties robustly disagree depends on our having assumed that each arrogantly believes that the other is subject to a normative requirement to bend to her wishes.

Intuitively, the Normative Disagreement Hypothesis ascribes the wrong prescriptive ‘force’ to the disagreeing thoughts of the parties in *Restaurant*. When the parties voice preferences, it sounds as though they are requesting that their preferred venue be chosen. Since requesting is a humble sort of action, the parties do not sound imperious and overweening when they cite their own self-regarding wants to explain their preferences. We fail to capture this quality of their disagreeing thoughts if we construe them as ought judgements, which require rather than request.

A defender of the Normative Disagreement Hypothesis might propose instead that we attribute a different kind of broadly normative belief to the parties, one less prescriptive in character. For instance, she might suggest the following as a more plausible characterization of the underlying disagreement in Stevenson’s case:

**Restaurant (Good)**

A: It would be good to go to a restaurant with music, because I feel like listening to music.

B: I disagree. It would not be good to go to a restaurant with music, because I don’t feel like listening to music.

This does not have the implausibly prescriptive flavour of the ought case. But we are not in a position to infer from their preference reports that the parties in Stevenson’s case hold the evaluative beliefs reported in *Restaurant (Good)*. This is true even if we assume that Stevenson’s parties very reliably form the desires that are appropriate given their evaluative judgements. Both parties in Stevenson’s case might think that going to a restaurant with music would be good (or not good), and both might think that going to one without music would be good (or not good); this is compatible with their having the
conflicting preferences they report, given that they reliably proportion their desires to their evaluative judgements. A party might think that neither option is good, or that both are good, and at the same time think that one is better than the other, and so prefer the former.

If we are in a position to infer evaluative judgements from the preferences Stevenson’s parties report, they will be judgements about which restaurant is the better choice. But in the decision-making situation that Stevenson describes, any such betterness judgement would have a clear prescriptive upshot. Thus, if we interpreted the parties as disagreeing about which restaurant would be the better choice, the grounds each gave for her preference would be evidence of a surprising arrogance. I conclude that the Normative Disagreement Hypothesis cannot plausibly explain all of our intuitions of robust disagreement.

1.1.5 Ridge’s ‘Disagreement in Prescription’

The observation we have made about the character of the disagreeing thoughts of the parties in Restaurant will be relevant to an evaluation of Michael Ridge’s theory of disagreement as ‘disagreement in prescription’. Two people disagree in prescription only if they are disposed to give contrary advice to someone under certain conditions:

**Disagreement in Prescription.** Two people (or two stages of the same person at different times), A and B, disagree in prescription about D’s φ-ing in C just in case in circumstances of honesty, full candour, and non-hypocrisy, A would advise φ-ing in C and B would advise ψ-ing in C, where φ-ing and ψ-ing are incompatible.70

To determine whether A and B disagree in prescription about D’s φ-ing, we are to ‘keep the states of mind of A and B fixed and ask, given those states of mind, what each of them would advise D to do if they had to advise D one way or the other, and moreover had to do so honestly, candidly, and without hypocrisy of any kind.’71 By ‘hypocrisy’, Ridge means advising actions that you would not perform yourself in the relevant circumstances.72

The scope of Ridge’s theory is ambitious: he apparently aims to give necessary and sufficient conditions for disagreement of all kinds.73 It is thus a theory of disagreement as ‘disagreement in prescription’, not a theory about a particular species

---

70 Ridge, Impassioned Belief, 187.
71 Ridge, Impassioned Belief, 187.
72 Ridge, Impassioned Belief, 186.
73 This is suggested by his remark in Ridge, Impassioned Belief, 169n.1.
of disagreement. With some modifications,⁷⁴ it purports to explain why incompatible (cognitive) beliefs put their possessors in disagreement. I will only consider its adequacy as a theory of non-cognitive disagreement, however. Specifically, I shall consider its adequacy as a theory of robust non-cognitive disagreement. Ridge’s theory of disagreement is not, of course, explicitly a theory of robust disagreement, but that doesn’t mean that it will not serve as one. Ridge does not want his theory to predict that cases like the flat cases are disagreements; he would class these as cases of ‘mere difference without disagreement’. So the set of non-cognitive disagreements that fall under his explanandum might be similar in membership to the set of robust non-cognitive disagreements.

Ridge’s theory has intuitively correct implications concerning most of the cases we have considered so far. It fairly plausibly predicts disagreement in Stevenson’s invitation case: Mr. and Mrs. A are plausibly disposed to advise incompatible options in the choice of whom to invite to their party. It does not predict disagreement in, for instance, Flat (Preferences), because, though each party wants the other to forbear from renting the flat, neither would honestly advise the other to forbear. A party might be disposed to dishonestly or ‘hypocritically’ advise the other to forbear, in the hope of eliminating competition, but disagreement in prescription depends only on dispositions to honestly and ‘non-hypocritically’ advise.

We find a pattern in the predictions of Ridge’s theory as it applies to candidate non-cognitive disagreements. It predicts disagreement when and only when it is plausible to suppose that the parties have conflicting relevant normative judgements. Mr. and Mrs. A seem to have conflicting judgements about whom they ought to invite; we are given no indication that each party in Flat (Preferences) has a judgement about which of them ought to get the flat, or about whether the other party ought to take it. The same is true, mutatis mutandis, for Retail and Rival Hostesses. This is to be expected of course: we only tend to have the intuition that someone would (honestly etc.) advise a person S to φ when we take it that she judgess that S ought to φ. So if Ridge’s theory is true, we should not expect to find any plausible cases of (robust) non-cognitive disagreement between people who do not plausibly share a relevant normative disagreement.

⁷⁴ Ridge, Impassioned Belief, 190.
As I observed in the last subsection, this prediction seems to be false. It is not plausible that the parties in Stevenson’s restaurant case are in normative disagreement, and yet they robustly disagree. More specifically, it is not plausible that either party is disposed to advise a choice of one dining venue rather than another, given the self-regarding grounds that she offers for her preference. Again, it is natural to think of the parties as requesting their preferred venues. Advising is quite different from requesting; it has a prescriptive flavour. Ridge’s theory thus seems to be false as a comprehensive theory of robust non-cognitive disagreement. At best it describes only true sufficient conditions for robust non-cognitive disagreement.

The moral philosopher who hopes to explain moral judgement as a species of non-cognitive disagreement will be more interested in knowing sufficient conditions for robust non-cognitive disagreement than in knowing necessary conditions. So Ridge’s theory may still be of interest. An evaluation of Ridge’s theory as an account of the sufficient conditions for robust non-cognitive disagreement is more than I can attempt here. Such an evaluation would need to take into account the success or failure of Ridge’s theory of disagreement in prescription as an explanation for our intuitions about doxastic disagreement, and several interesting arguments that Ridge offers in defence of the theory. But the failure of Ridge’s theory to provide true necessary conditions gives us enough reason to be dissatisfied with it to warrant a search for alternatives.

1.1.6 A Guiding Thought: The Reason-to-Consider Hypothesis

I doubt that we can make progress in explaining the robustness of the disagreements Stevenson describes merely by hypothesizing that we interpret them as involving ‘clashing’ attitudes of one type rather than another. Something is missing in these accounts. And Stevenson’s restaurant case suggests that the missing ingredient isn’t normative judgement or dispositions to advise.

---

75 These are listed and partly summarized in Ridge, Impassioned Belief, 188-191.
76 Note a disappointing feature of Ridge’s account for the non-cognitivist. Someone advancing a non-cognitivist theory of advice might have hoped to have supported this theory by showing that it explains why people who give conflicting advice disagree. This would probably mean showing that the non-cognitive attitudes that are expressed in the giving of advice are capable of disagreeing, by showing that this is entailed by some independently plausible theory of non-cognitive disagreement. If Ridge is correct, this potential avenue of support is lost, for the truth of his theory would make the plausibility of the claim that the relevant attitudes put their possessors in disagreement dependent on the plausibility of the claim that they dispose their possessors to give conflicting advice.
I will now offer a suggestion which I find promising though vague, and which may help to focus our investigation. There is some intuitive plausibility in the idea that two people whose thoughts are in disagreement must have some reason (perhaps very weak) to hear and give consideration to each other’s contrary thoughts, or else they would not robustly disagree. Call this the Reason-to-Consider Hypothesis. It has prima facie plausibility as an explanation of the robustness of the disagreements Stevenson describes. The parties to Stevenson’s restaurant case have reason to hear and consider each other’s preferences regarding which restaurant they will dine in—each wants the other to remain favourably disposed toward dining with her, and this is more likely to be achieved if she maintains a polite level of attentiveness to the other’s wishes, and a limited willingness to accommodate them. There are various reasons why Mr. and Mrs. A might want to accommodate each other’s preferences, and hence to hear and consider them. They may desire to do so because they want each other to remain favourably disposed toward their shared project of holding a dinner party, or because a reciprocal willingness to accommodate each other’s preferences makes living together more tolerable, or because social norms require husbands and wives to give regard to each other’s wishes. Or perhaps mutual affection, of the sort we expect to find between husband and wife, prompts this sort of attentiveness.

Charitably interpreted, the Reason-to-Consider Hypothesis makes no false predictions about the other cases we have considered. The parties in Flat (Preferences) evidently have no reasons to hear and give consideration to each other’s preferences. Customer in Retail has no reason to give consideration to Shopkeeper’s preference when choosing whether or not to buy the expensive coat. For strategic reasons, each party in Rival Hostesses might be interested to know about the other’s preference, for this would be evidence that the other was working to frustrate her own. But this intuitively would not be a reason for the hostesses to hear and give consideration to each other’s preferences.

Interestingly, we find some support for the Reason-to-Consider Hypothesis when we consider disagreement in belief. Imagine the following exchange between a medical expert, Dr. Expert, and a layperson:

**Vaccine**

*Dr. Expert:* The measles-mumps-rubella vaccine does not cause autism.

*Mr. Layman:* I believe it does. We disagree.
Dr. Expert: Disagree? How presumptuous! Your medical opinions are beneath my notice.

Arrogant it may be, but Expert’s reply makes sense. Mr. Layman’s disagreement claim does sound presumptuous, which would be odd if all that were required for doxastic disagreement were contradictory beliefs. It seems that Expert could correctly reject Layman’s claim that they disagree, on the grounds that the latter’s belief does not merit her attention (assuming this was the case). Since Expert’s reply is ‘conciliatory’ in my stipulated sense, Vaccine seems to be a case of non-robust doxastic disagreement. The case thus supports, or at least is in harmony with, the Reason-to-Consider Hypothesis.

Consider a second case. Liar knows that ~p but says that p, causing Misled to believe that p. Liar and Misled now have conflicting beliefs about whether p, but it is a bit strange to say that they disagree. (It would certainly sound odd for Liar to say of Misled, ‘She disagrees with me about whether p.’) And Liar has no reason to hear and give consideration to Misled’s opinion that p, for Liar has seen to it that it is false. Finally, as Ridge observes, we intuitively don’t disagree with non-human animals whose beliefs differ from ours; perhaps this is because animals and humans can’t hear and give consideration to one another’s thoughts, or because we take animals’ ‘opinions’ to be ‘beneath our notice’. It is noteworthy that, as in non-robust non-cognitive disagreement cases, there is intuitively no ‘matter to be settled’, between Expert and Layman, between Liar and Misled, or between a cat and its owner when their beliefs are at odds.

Whether or not we take into consideration these observations about doxastic disagreement, we have grounds to think that a version of the Reason-to-Consider Hypothesis can explain the phenomenon of robust non-cognitive disagreement. But I have only sketched it so far, and it needs to be stated more precisely before it can be properly evaluated. It raises several questions: (1) What does it mean to ‘hear and give consideration’ to a preference? (2) If someone is to be in robust non-cognitive disagreement with someone else’s preference, is it enough for her to have any kind of reason to hear and give consideration to that preference? (3) Does it make robust disagreement a fundamentally normative notion, or is the phrase ‘have some reason’, where it appears in the statement of the hypothesis, just shorthand for ‘be related in such-and-such a way to some member of a certain set of non-normative facts’, where

---

77 Ridge, ‘Disagreement’, e.g. 54.
these are facts that we would normally judge to be reasons in the relevant context? (4) Does (or should) the hypothesis provide sufficient or necessary conditions for an interpersonal clash of preferences to be a robust disagreement (or both)?

The remainder of this chapter will be devoted to identifying a version of the Reason-to-Consider Hypothesis that answers these questions and explains the phenomena.

1.1.7 The Coordination Hypothesis

The prospective diners in Stevenson’s restaurant case face a ‘bargaining problem’. They have a common goal of dining together which they would best serve by coordinating their actions, and they apparently have indeed decided thus to coordinate their actions. But there are different ways in which they could coordinate to realize their shared end, and they have preferences for mutually incompatible courses of coordinated action. It is these preferences that put them in disagreement. And it might be hypothesized that they only put the parties in robust disagreement because they are preferences for mutually incompatible courses of coordinated action, in a context where the parties have together decided to coordinate their actions. We might say that such preferences in such a context suffice for robust non-cognitive disagreement. Call this the Coordination Hypothesis.78

The Coordination Hypothesis can be seen as a way of giving substance and detail to the Reason-to-Consider Hypothesis, as it applies to non-cognitive disagreement. It allows us to answer some of the questions that my sketch of the latter hypothesis raised. (1) To ‘hear and give consideration to’ a preference, one must be a party to an agreement to coordinate actions in pursuit of a common end. Hearing and giving consideration to a preference involves, we might say, hearing the preferences of the other parties for different courses of coordinated action, and deciding whether or not they require you to adjust your own demands, or the outcome of the negotiation that you are aiming for. It is plausible that any party to such an agreement has reason to hear and give consideration to the preferences of the other parties, when deciding on a coordinated course of action. She has this reason because she has a goal that would be furthered by cooperating with the others, and cooperation would potentially break down if she were not responsive in some degree to the other parties’ relevant preferences. (2)

78 Cf. Gibbard’s suggestion that it is an important feature of Stevenson’s examples that the parties disagree about ‘how to act jointly’. Gibbard, Thinking How to Live, 270.
Only reasons of this sort make a difference to the robustness of disagreements in preference (for all the Coordination Hypothesis says). (3) However there is no real need to rely on the normative notion of a reason here; fundamentally, it is being a party to an agreement of the relevant sort that matters. (4) At this stage I have only presented the coordination hypothesis as providing a sufficient condition for robust non-cognitive disagreement.

The Coordination Hypothesis seems to make only true predictions about the cases we have considered. It would be natural to think of Mr. and Mrs. A in Stevenson’s invitation case as facing a coordination problem, just as the parties in the restaurant case do. They have a common goal of hosting a successful party which they would best serve by coordinating their actions, and they apparently have indeed decided thus to coordinate their actions. They disagree (on at least some plausible interpretations of the case) about which of two incompatible courses of coordinated action to pursue, each of which involves inviting a different guest. The Coordination Hypothesis does not falsely predict disagreement in Flat (Preferences), Retail, or Rival Hostesses. None of the parties in these cases appear to have decided to coordinate their actions, nor do they evidently have any common goals that would be served by cooperation.

Despite these merits, the Coordination Hypothesis cannot explain all of our intuitions about robust non-cognitive disagreement. There are cases of robust non-cognitive disagreement in which the parties cannot be assumed to have decided to coordinate their actions. Consider a third case from Stevenson:

John’s mother is concerned about the dangers of playing football, and doesn’t want him to play. John, even though he agrees (in belief) about the dangers, wants to play anyhow. … [T]hey disagree.79

I did not quote this case earlier because, although it is plausibly a case of disagreement, I don’t think Stevenson’s description is detailed enough to establish that the disagreement is robust. But imagine the following two exchanges, the first between John and his mother, and the second between two eavesdroppers who have overheard this exchange:

**Football (Mother and Son)**

_Mother:_ Why are you wearing football shoes in here? You know I don’t approve of shoes in the house. And you’re leaving dirt everywhere!

_John:_ Sorry! I’ll take them off.

Mother: I’d prefer that you didn’t play football anyway. It’s dangerous.
John: I know, but I want to play. I love football.
Eavesdropper A: [To Eavesdropper B] Sounds like a disagreement.
Eavesdropper B: [To Eavesdropper A] I don’t think they disagree. They just prefer different things.

(The opening lines of this dialogue are meant to establish that a certain sort of relationship exists between John and his mother. I included the Eavesdroppers because putting the conciliatory claim in the mouth of John or his mother would have sounded odd.)

Conciliation seems inappropriate here; John and his mother disagree robustly about whether John is to play football; there is intuitively an unresolved issue, a ‘matter to be settled’, between the two. And John and his mother do not face a coordination problem. Their disagreement cannot be about how to coordinate their actions, for, of the two parties, only the son is contemplating performing an action. It would not be unreasonable to assume that the parties have a common goal—a common desire to see their family flourish in one way or another. This would be a fairly normal goal for a mother and son to share. But it is far from clear that they are disagreeing about ways to achieve such a goal. John seems to see his playing football as a means of achieving a more private end, viz., his own enjoyment. The coordination hypothesis thus does not well explain our judgements about this case.

1.2 The Entrusted Concerns Hypothesis

The coordination hypothesis seems to be on the right track when it identifies the relationship between the parties to a robust non-cognitive disagreement as a determinant of its robustness, but it seems to give too narrow a specification of the sorts of relationships that suffice to make an opposition of preferences a non-cognitive disagreement. The robustness of John’s disagreement with his mother appears to depend on the nature or their relationship.

Consider a similar exchange between John and Well-Wisher, a benevolent stranger who wants things to go well for John, and who is worried that John might get hurt playing football.

Football (John and Well-Wisher)
John: Hi. I’m John. What’s up?
**Well-Wisher:** I work for the Gas Company, and while I was checking your meter, I overheard you and your mother talking about football. And like your mother, I would prefer that you didn’t play. It’s dangerous.

**John:** Well, as you may have heard me tell my mother, I know it’s dangerous, but I still want to play—I love football.

**Eavesdropper A:** [To Eavesdropper B] Sounds like a disagreement.

**Eavesdropper B:** [To Eavesdropper A] I don’t think they disagree. They just prefer different things.

(The opening introductions are supposed to make vivid that the speakers are perfect strangers.) Eavesdropper B’s conciliatory claim is appropriate here. There is no ‘matter to be settled’ between John and Well-Wisher. John might indulge Well-Wisher by treating her preference as deserving his consideration, and then we might be inclined to think that they disagree; but presumably he is indifferent to it. The only important difference between this clash of preferences and that between John and his mother is a difference in the relationship between the parties.

1.2.1 **Relevant Features of John’s Relationship with his Mother**

In what follows I offer a conjecture as to the features of the relationship between John and his mother that explain the robustness of their disagreement. John’s mother evidently wants John to be safe. She also has power over John, so she could probably restrict his freedom to engage in dangerous activities to a greater extent than she in fact does if she wished. But, since she also presumably wants John to be happy, to enjoy a degree of freedom, and to take responsibility for his own actions, and since personally managing someone else’s behaviour can be a chore, it is plausible to suppose that she has instead partly *entrusted* or delegated her concern for John’s safety to John himself. She plausibly trusts John to have a special concern for his own safety, over and above the concern he would have for his safety if left to his own devices.

‘Concerns’ need not be decisive in decision-making, and John plausibly has some discretion to resist his mother’s desire. He may weigh his special concern to accommodate his mother’s desire for his safety against his other more private concerns. He might thus knowingly risk bodily injury, in some circumstances and to some extent, and thereby risk frustrating his mother’s desire, without violating her trust. The trust relationship I have described would suit John, because of the greater freedom it would afford him, so we may assume that he accepts his mother’s trust and is committed to living up to it.
John’s concern to accommodate his mother’s desire that he be safe is plausibly just one of several such concerns his mother has entrusted to him. She probably has other desires concerning John’s welfare: desires for John to be healthy, to be educated and accomplished, and to have friends. She also probably has desires that are not focused on John in particular, but which she will want John to have some concern to accommodate, such as a desire to improve her family’s welfare. If this is right, it is plausible to suppose that she entrusts to John a concern to promote these ends.

Moreover, she will want the special weight John gives to promoting these various ends in his decision-making to be proportionate to the strength of her corresponding desires, and this will be reflected in their trust relationship. If John must, in some instance, choose between health and accomplishment, and his mother has a preference, based purely on her desires that he be healthy and accomplished, that he choose the health-promoting option, this is the option his trust relationship will direct him toward. It will do so because the weight that John is to give to health and accommodation depends on the strength of his mother’s desires that he be healthy and accomplished respectively, and her preference is determined by the relative strength of these desires. So though it is correct to say that John has an entrusted concern to promote the satisfaction of some of his mother’s desires, it is also correct, and in some ways more informative, to say that he has an entrusted concern to accommodate those of her preferences that are based purely on those desires.

John’s discretion to resist accommodating his mother’s preferences also plausibly depends on the motivations he would have for doing so. John might have discretion to continue playing football, in defiance of his mother’s preference, if he did so because he enjoys football and desires enjoyment, or because playing football keeps him fit and he wants to be fit. But he might not if he merely did so because all the popular kids play football and he wants to emulate the popular kids, or because the coach gives illegal steroids to team members and he wants to take steroids.

It is not hard to imagine how the motivations of both John and his mother might have shaped the trust relationship between them, in such a way as to determine which of their desires John was expected to accommodate, and which he wasn’t. The power John’s mother has to induce John to make a commitment to accommodate her desires has limits, and she will thus want to make the demanded commitment relatively lenient, so that John will not find it too irksome. Given that it must be lenient and not too demanding, she will want it to be demanding only in the respects that matter to her the
most, and lenient in those respects that she finds most tolerable. She will want John to be committed to accommodating her desires that he be safe and healthy, that he should become educated and accomplished, and that he should promote their family’s welfare. But she will be willing to sacrifice her capacity to demand that he give consideration to most of her other desires. Likewise she will want John to have considerable discretion to indulge his more innocent desires, e.g., to have fun (both to make membership of the trust relationship attractive to John, and because she herself presumably wants him to be happy). But she will not want him to have discretion to resist accommodating her preferences if it is so that he may indulge those of his desires (e.g., his desire to emulate the popular kids, his desire to take steroids) such that his attempts to satisfy them might lead him to develop traits, motivations, and relationships inimical to his future success and happiness.

I am going to argue that it is John’s resistance to his mother’s preference, in a context in which she trusts him to give consideration to this preference, that explains why the two robustly disagree in Football (Mother and Son). There is a gap between what his mother voices a preference that he should do and what he proposes to do, and this places strain on her trust. John’s resistance does not amount to an outright breach of her trust, but only because he resists her preference in accordance with a preference of his own that is based on motivations of the sort I have described. The strain he puts on her trust by disagreeing with her creates, I suggest, a sense of tension and conflict between the pair such as we expect to find between the parties to a disagreement.

More generally, I will argue that robust non-cognitive disagreement requires the existence of a trust relationship between the parties similar to that which we find between John and his mother. It involves a clash of preferences between two parties, concerning what one of those parties is to do. The preference belonging to the non-agent party is based on desires that she has entrusted the agent to accommodate; the preference belonging to the agent is based on desires that licence him to resist accommodating the former desires without a violation of trust. I will call this view the Entrusted Concerns Hypothesis.

1.2.2 Entrusted Concern Relationships

I’ll now introduce some terms and conventions with which to describe the features of the relationships that make robust non-cognitive disagreements possible, and at the same time try to give a better idea of how these relationships would work. This
discussion is more elaborate than it needs to be for my present purpose of explaining robust non-cognitive disagreement. But I will be talking about these relationships and their features throughout this dissertation, and all of the details I provide here will prove important at some point.

I will call desires such as those that John’s mother trusts John to accommodate, and those that licence him to resist doing so, \textit{weighted desires}. I call them this, not because they are given a weight that differs from the usual weight a desire of the same strength has in practical deliberation, but simply because they are, unlike other desires, allowed and required to weigh in the deliberations of people who occupy positions such as John occupies, in trust relationships such as John shares with his mother. It is partly in order to establish that the preferences that they voice are based on weighted desires that the parties to robust non-cognitive disagreements often report the grounds for their preferences.

I call relationships of the sort that we find between John and his mother \textit{entrusted concern relationships}. In such relationships the party who has entrusted to the other party a concern to accommodate her preferences (e.g. John’s mother) is the \textit{entrusting party}; the other party (e.g. John) is the \textit{entrusted party}. (By default, I will refer to an entrusting party as ‘she’ and an entrusted party as ‘he’.) Each party in an entrusted concern relationship is the other party’s ‘partner’. The weighted desires of parties to an entrusted concern relationship are desires that fall under \textit{weighted desire-types}, which are specified by the terms of that relationship. I will call those preferences of an entrusting party that are based purely on her weighted desires—i.e. those preferences she would have if all non-weighted desires were omitted from her set of basic desires—\textit{accommodation-obliging preferences}; and those of entrusted parties \textit{discretion-licencing preferences}. These are preferences that the parties are entitled, according to the terms of their trust relationships, to \textit{voice} in the entrusted party’s decision. Thus I will sometimes call accommodation-obliging and discretion-licencing preferences \textit{voiceable preferences}.

Just as it is possible to report a preference insincerely by knowingly reporting a preference that one lacks, it is possible to \textit{voice} a preference insincerely. However, there is an unsurprising difference between insincerely reporting and insincerely voicing a preference. When a person voices a preference that \( p \), she conventionally indicates that, qua party to some contextually salient entrusted concern relationship, she has a \textit{voiceable} preference that \( p \), i.e., one that is based on only a weighted subset of her
desires. To insincerely voice a preference that \( p \) is thus to (knowingly etc.) indicate that one has a voiceable preference that \( p \) when one does not, even if one has a preference of some other kind (e.g. an all-things-considered preference) that \( p \)—that is, even if one has a preference that \( p \) that one might sincerely report.

When describing entrusted concern relationships it is very natural to use normative terms, and I will not try to avoid doing so. Thus, for instance, I shall say that entrusted parties are subject to trust-based obligations, and that both entrusted and entrusting parties enjoy trust-based entitlements. I don’t regard this usage as stipulative or loose; these are obligations and entitlements of the familiar sort that conventions and understandings can give rise to. I assume that they are not ‘genuinely’ normative in the elusive sense that moral obligations are; and that they are susceptible to a comparatively straightforward naturalistic reduction. The viability of my metaethical project depends on this.

I have said that John strains his mother’s trust because he ‘proposes’ not to act in accordance with her accommodation-obliging preference. I will henceforth say instead that the strain an entrusted party puts on his partner’s trust results from his failure to yield to her accommodation-obliging preference. If John gave his mother his assurance that he would not play football, he would yield to her preference, and they would no longer disagree. He would, I suggest, also yield if he actually did as his mother preferred, without having given his assurance first. Entrusting parties may also yield. If John’s mother gave John her blessing to play football, she would yield to his discretion-licencing preference. They would no longer disagree, and John could choose to play football without straining their trust relationship.

Robust non-cognitive disagreement only exists when neither party has yielded. But I don’t think that yielding ends disagreement, save perhaps in odd cases. Rather, disagreement typically ends before the act of yielding has been performed, as soon as one party forms an intention to yield. Suppose that John’s mother formed an intention to yield.

---

80 Here are some properties that these trust-based obligations have, but whose possession by moral (and other ‘genuinely’ normative) obligations is controversial; we can sensibly ask whether we ought to act on these trust-based obligations; their practical ‘authority’ over us is contingent—usually on our desire to remain parties to the agreements (etc.) that give rise to them; we almost always honour these obligations for instrumental reasons, or because doing so is habitual, or, more strongly, because we are ‘in their grip’, or have ‘internalized’ them, in the senses that Gibbard describes (Wise Choices, 60-61, 68-71); two people with the same beliefs about the non-normative facts (on a suitably narrow construal of ‘non-normative’) cannot hold quite different beliefs about these trust-based obligations unless at least one of them is confused or irrational.
yield to John’s preference for playing football; it seems that there would no longer be any disagreement—no ‘matter to be settled’—between John and his mother, even if John still thought there were. Thus our theory of robust non-cognitive disagreement needs to make mention of yielding intentions as distinct determinants of disagreement.

Yielding need involve no change in preferences among options. If a party to an erstwhile disagreement intends to yield to the other party’s preference for some option, this does not necessarily mean she has come to prefer that option. The party’s accommodation-obliging or discretion-licencing preference may be unchanged. Indeed, in the case of yielding by entrusting parties, the party’s ordinary, ‘non-voiceable’, all-things-considered preferences between the options may not have changed. John’s mother might intend to yield to John’s preference for playing football because she doesn’t want to make a fuss, even though her desires (weighted and otherwise) concerning his playing football remain unaltered.

I find it very natural to think of the stance of refusing to yield to the selection of a certain option, when adopted by an entrusting party, as a stance of demanding that the alternative option be chosen. So henceforth, if I talk of entrusting parties ‘demanding’ actions, or describe an entrusted party’s stance of not yielding as one of ‘resisting demands’, this is what I mean by the word. Note that the ordinary meaning of ‘demand’ might give this usage misleading resonances, for an entrusting party’s refusal to yield may place only very gentle trust-based ‘pressures’ on the entrusted party to accommodate her preference.

An entrusted party strains his partner’s trust by refusing to yield and relieves strain by yielding. If he places too much strain on her trust over a series of decisions, he will eventually violate it. Thus, if the entrusted party wants to honour his commitment and so preserve his trust relationship, he will generally want to avoid ‘accumulating’ too much strain; so he will want to yield from time to time. (I make a proposal concerning entrusting parties’ reasons for yielding in Appendix A.)

Having introduced the notion of yielding, I’d better clarify my earlier characterization of voiceable preferences. Once an entrusting party has yielded to her partner’s preference in some decision, she is no longer entitled to voice a preference in that decision, and her preference is no longer one that her partner is obliged to accommodate. Despite this, I have found it convenient to keep referring to the preferences she is no longer entitled to voice as her ‘voiceable preferences’ and her ‘accommodation-obliging preferences’. Likewise, I have found it convenient to keep
referring to an entrusted party’s preferences as his ‘voiceable’ or ‘discretion-licencing’ preferences, even if he has yielded, in the sense of giving his assurance that he will acquiesce. Thus, on my usage, a preference’s status as voiceable, accommodation-obliging, or discretion-licencing does not depend on whether or not its possessor has yielded.

A final hypothesis. I propose that weighted desires must be basic desires—roughly, desires that do not derive from other desires in combination with beliefs about what would satisfy the latter desires. It would be counterintuitive to suppose that, in the entrusted concern relationship shared by John and his mother, John might be obliged to accommodate his mother’s desire that he be safe, if she only wants him to be safe so that she can win a bet. The hypothesis that weighted desires must be basic desires explains how this could be ruled out: the terms of the trust relationship between John and his mother might specify that John’s mother’s weighted desires are for things like promoting her family’s welfare, and since the weighted desires would be basic, if she had desires for promoting her family’s welfare (etc.) that derived from her desire to win a bet, these would not be weighted desires. If the hypothesis were not true, then, even if John’s mother’s weighted desires were only for things like promoting her family’s welfare, it would be possible that John might be obliged to accommodate one of these desires if it was derived from a desire to win a bet. For instance, if the hypothesis were not true, the following counterintuitive state of affairs would be possible: (i) John is obliged to stop playing football because football is dangerous, because an injury to John would diminish his family’s welfare, and because his mother has a weighted desire to improve her family’s welfare; and (ii) John’s mother has this desire merely because she wants to win a bet about whether or not she can improve her family’s welfare.

Admittedly, I’m not sure that something weaker than the hypothesis that weighted desires must be basic desires could not explain these intuitions. This feature of my characterization of entrusted concern relationships should be regarded as relatively conjectural.

I warn the reader that I shall not be very careful, when attributing token weighted desires to parties, about establishing that they really are basic (and hence possibly weighted) desires, instead of being derived from other genuinely weighted desires. There are some derived (i.e. non-basic) desires that are sufficiently robust with respect to ordinary changes in their possessors’ beliefs that they behave a lot like basic desires. It will sometimes be natural to treat such a desire as a weighted desire, or at least to call
a desire a weighted desire without first eliminating the possibility that it is one of these robust derived desires. If John’s mother has a desire that John should be safe, this desire might derive from a (weighted) desire that John should be happy and successful; nonetheless, I may, speaking loosely, call the former desire a weighted desire without worrying about this possibility, or bothering to rule it out.

1.2.3 A Statement of the Entrusted Concerns Hypothesis

In a moment I shall present what I take to be an approximation of necessary and sufficient conditions for robust non-cognitive disagreement. But first, I shall introduce the notion of a W-based preference, and a definition of ‘entrusted concern relationship’.

**W-based preferences.** For all S, all W, and all propositions p, S has a W-based preference that p iff W is a set of desire-types (typed solely by content, not strength, or any other property), and S would prefer that p if all of those basic desires of S that did not fall within the types in W were omitted from her set of basic desires.

**Entrusted concern relationship.** For all trust relationships E, all A, and all B, E is an entrusted concern relationship in which A is the entrusting party and B the entrusted party iff there is some set of types of basic desires W (typed solely by content, not strength, or any other property), some set of types of basic desires X (also typed solely by content), and some class of decisions D such that

(i) the terms of E require B to accommodate all and only A’s W-based preferences in his D-decisions (unless A consents to his doing otherwise⁸¹), while allowing B some discretion to resist doing so when this would frustrate his X-based preferences;

(ii) A trusts B to comply with the terms of E; and

(iii) B is committed to living up to this trust.

Now the hypothesis:

**Entrusted Concerns Hypothesis.** For all persons A, all persons B, and all φ, A and B are in robust non-cognitive disagreement about whether B is to φ if and only if there is some E, some set of types of desires W (typed solely by content, not strength, or any other property), some set of types of basic desires X (also typed solely by content), some class of decisions D, and some ψ such that:

(i) E is an entrusted concern relationship in which A is the entrusting party and B the entrusted party;

(ii) E obliges B to accommodate A’s W-based preferences in all D-decisions B faces, while allowing B some discretion to resist accommodating A’s W-based preferences in D-decisions when and only when accommodating them would frustrate B’s X-based preferences;

(iii) ψ and not-ψ are jointly exhaustive mutually incompatible options available to B;

(iv) B’s choice between ψ and not-ψ is a D-decision;

---

⁸¹ If B’s X-based preferences oppose A’s W-based preferences, such consent will amount to yielding. Otherwise it will involve ‘waiving’, a notion I introduce in 2.2.1.
The characterizations of entrusted concern relationships and the Entrusted Concerns Hypothesis that I have presented thus far are somewhat lacking in detail, and are in some respects incorrect. Supplying some missing details and remedying the errors will prove important to my argument in this dissertation. But I have opted to postpone the necessary elaborations and corrections until I am ready to explain why they are of interest. We can make good progress in evaluating the Entrusted Concerns Hypothesis by considering the version that I have presented in this chapter, despite its incompleteness.

Like the Coordination Hypothesis, the Entrusted Concerns Hypothesis can be regarded as a way of cashing out the Reason-to-Consider Hypothesis. Let me briefly list the answers it offers to the questions that the Reason-to-Consider Hypothesis raises. (1) According to the Entrusted Concerns Hypothesis, hearing and giving consideration to the ‘disagreeing thoughts’ of someone with whom one is in robust non-cognitive disagreement involves being responsive to trust-based demands and entitlements. Entrusted party B has reason to hear and give consideration to a preference voiced by his partner A because he may be obliged to accommodate it. If B does not acquiesce, A has reason to hear and consider B’s voiced preference because it is this alone that could establish that B is licenced to resist her preference, and so would not violate her trust by doing so (assuming that B has not strained A’s trust to breaking point). Each, moreover, has reason to scrutinize the grounds the other offers for her preference, to determine how plausible it is that she truly has the accommodation-obliging or discretion-licencing preference that she purports to have. (2) These are the only sorts of reasons to ‘hear and give consideration’ that the parties may possess that are relevant to an assessment of whether or not they disagree robustly.

(3) The Entrusted Concerns Hypothesis does not seem to make robust non-cognitive disagreement judgements dependent on ‘genuinely’ normative judgements. The reasons the parties to such disagreements have to hear and give consideration to

---

82 They appear in 2.2.1 (‘waiving’), 3.1.4 (‘discursive entitlements’), 3.3.3 (more on waiving, ‘admissible grounds’), 5.4.1 (‘objective trust’), 6.2.1 (fundamental and derived entitlements), and 8.2.1 (the Disparity Hypothesis); and Appendices A (‘debt’), B (‘conflict excuses’), and E (‘waiving’ again).
each other’s preferences are of an instrumental sort. The entrusted party’s reasons are explained by his desire to avoid straining and violating his partner’s trust, something he cannot do unless he heeds her voiced preferences. The entrusting party’s reasons are explained by her weighted desires, and the fact that she can expect these to be better satisfied by the entrusted party’s immediate decision if he does not enjoy discretion to resist her preference. We could apparently dispense with talk of reasons, and posit that it is this combination of desires and trust-based entitlements that explains why the parties disagree. So the Entrusted Concerns Hypothesis does not turn ascriptions of robust non-cognitive disagreement into normative judgements. (4) I have boldly formulated the Entrusted Concerns Hypothesis as a biconditional, giving necessary and sufficient conditions for robust disagreement between mutually unsatisfiable preferences. We shall consider later whether I have been too bold in this regard.

In the next two subsections I will evaluate the Entrusted Concerns Hypothesis. I will first consider whether it makes the right predictions about the cases I have already considered, and a few variations on these cases. I then consider whether disagreement in imperatives might pose a challenge to the hypothesis.

1.2.4 Predictions of the Entrusted Concerns Hypothesis in Other Cases

The Entrusted Concerns Hypothesis can, unlike the Coordination Hypothesis, explain our intuitions about Football (Mother and Son). But I haven’t yet shown that it explains our intuitions about the other cases we have considered. It is easy to see that the Entrusted Concerns Hypothesis makes the right predictions about Flat (Preferences), Retail, and Football (John and Well-Wisher). These are not cases of robust non-cognitive disagreement, and in none of them is it plausible that an appropriate trust relationship exists between the parties. It is less obvious that the hypothesis can explain our intuitions about Stevenson’s other cases.

Among Stevenson’s cases, showing that the football case meets conditions (i)-(vii) of the Entrusted Concerns Hypothesis is a relatively simple business. For, in the

---

83 Two caveats. First, the Entrusted Concerns Hypothesis does require that such ascriptions involve judgements of about the trust-based obligations and entitlements of the parties, and I am not able to completely rule out the possibility that these are ‘genuinely’ normative judgements. Second, it is still possible, that some version of the Reason-to-Consider Hypothesis is what explains why the disagreements predicted by the Entrusted Concerns Hypothesis seem to be disagreements, and that this version of the Reasons-to-Consider Hypothesis does turn disagreement judgements into normative judgements.

84 The hypothesis makes no predictions about Rival Hostesses, for the clash of preferences in that case concerns the actions of a third party, not one of the parties to the disagreement.
football case, it is clear who must be the entrusting party and who the entrusted party. In the invitation and restaurant cases this is not clear. I propose that in these cases the parties share *two* entrusted concern relationships; each party stands in a relevant ‘entrusting’ relation to the other. They disagree about two pairs of options: whether or not party X is to assent to party Y’s preferred course of action, and whether or not Y is to assent to X’s preferred course of action.

I will start by testing the hypothesis against Stevenson’s invitation case. It is natural to suppose that Mr. and Mrs. A have entrusted various concerns to each other, because this is something husbands and wives generally do, and because it presumably makes living together more tolerable. Mr. A and Mrs. A plausibly entrust to each other concerns for their health and welfare, and for the prosperity of their household. It may not seem obvious that these are concerns to accommodate desires—our evidence for this is that it would seem natural for them to assert their demands as parties to these trust relationships by voicing wants and preferences. It is also plausible to suppose that they have entrusted to each other concerns to accommodate their more selfish desires, concerning choices affecting them both (e.g. choices about what furniture to buy). Thus it would not be far-fetched to suppose that Mr. A trusts Mrs. A to give weight to his desire that they should honour their old friendships in her decisions, or that Mrs. A trusts Mr. A to give weight to her desire that their social status be improved in his. Given the Entrusted Concerns Hypothesis, it is easy to see how each of these trust relationships could make a robust disagreement of their clashing preferences in *Invitation*.

When the prospective diners in Stevenson’s restaurant case agree to dine together, we plausibly assume that they entrust certain concerns to each other, and make corresponding commitments to give weight to these concerns in their decision-making. Prospective diner A plausibly entrusts to prospective diner B a concern to make their dinner pleasant and convenient for A, insofar as A desires such attentiveness from B; and vice versa. In other words, they trust each other to be considerate. B’s entrusted concern to make A’s dining experience a pleasant one gives her reason to listen to A’s preference to go to a restaurant with music (say), and to give it special weight when deciding which restaurant to agree to go to. This entrusted concern is, however, lenient, and permits B to give weight to her own preference not to hear music. The Entrusted Concerns Hypothesis thus plausibly predicts that A and B robustly disagree in preference when they voice mutually unsatisfiable preferences about where to eat. (Of
course all that I have just said regarding B’s relationship to A also applies in the other direction, for A’s preference and her entrusted concern mirror B’s in all relevant respects. But the hypothesis predicts robust disagreement even when this fact is ignored.)

Could we think up versions of Stevenson’s restaurant and invitation cases that elicited intuitions of disagreement, but whose features discouraged us from attributing an entrusted concern relationship to the parties? We haven’t yet considered any cases in which two people who have little reason to trust each other nonetheless have strong motivation to reach an agreement on some coordinated course of action. Suppose that A is a blackmailer, and wants to meet B at a restaurant that she owns, late at night, so that she can exchange her scandalous photographs of B for money. B wants to make the exchange at a more public venue, perhaps because he fears that A will rob him of his money without giving him the photographs. We should not expect A and B to have much trust for each other. A and B do not agree about where to make the trade, and they cannot easily reach agreement. But do they disagree? I find it odd to suppose this; intuitively, the supposition represents their relationship as being implausibly civilized and respectful.

What happens if we dramatize their clash of preferences in a dialogue, as I did with Stevenson’s cases? It is in fact hard to do this without giving the case distractingly odd features which would interfere with any intuitions of disagreement it might otherwise have elicited. We can easily imagine A and B demanding to meet at different places, and being unable to reach agreement. Consider:

Restaurant (Blackmail I)

A: Meet me at my restaurant after closing time, and we’ll make the trade.
B: I don’t think so. It’s got to be a public place, somewhere neutral. Let’s meet at Minims at seven.
A: This isn’t a negotiation. Come to my restaurant, or I send the pictures to the press.
B: No way. We disagree.

The exchange of demands, ultimatums, and refusals here seems natural enough, and I think that B’s disagreement claim seems odd.

But suppose instead that A and B have the following exchange:

Restaurant (Blackmail II)

A: Meet me and we’ll make the trade. I’d prefer that we met at my restaurant after closing time.
B: I’d rather we met in a public place, somewhere neutral. How about Minims at seven?
A: I’d really prefer that we met at my restaurant.
B: We disagree.

This whole exchange seems strange given its context. Why would the parties be reporting and listening to each other’s preferences as if they were concerned to accommodate them? Admittedly B’s disagreement claim is not the strangest feature of the case, and perhaps we lack a clear intuition that A and B don’t disagree. But this is, I suggest, partly because the general strangeness of the case is distracting, and partly because A’s and B’s exchange of preference reports gives us some evidence that, bizarrely, each party is concerned to accommodate the other’s opposing preferences. I can’t claim that Restaurant (Blackmail II) gives resounding support to the Entrusted Concerns Hypothesis, but I don’t think it undermines it either.85

1.2.5 Are Cases of Imperative Disagreement Counterexamples to the Hypothesis?

Cases in which two people utter conflicting imperative sentences can elicit strong intuitions of disagreement. Consider the following.86

Drinks
Cody: Let’s get a coffee.
Sally: No, let’s get a beer.
Cody: No, let’s get a coffee.
Sally: We disagree.
Cody: I don’t think we disagree.

Cody’s conciliatory claim seems inappropriate. Sally and Cody seem to be in a robust disagreement, and it might be a non-cognitive one, for the parties do not (in the initial disagreement about what to get) express disagreeing beliefs. Indeed it is not absolutely clear that our intuition of disagreement is explained our attributing disagreeing thoughts.

85 Ridge describes a variation on Stevenson’s restaurant case (as a counterexample to Stevenson’s conditions for disagreement in attitude, as Ridge interprets them) in which “[t]he two people hate each other, but have to eat together.’ Partly out of spite, each party X wants the other party Y to go Y’s dispreferred restaurant. Neither will acquiesce to the other’s wishes. Ridge thinks that this is a case of disagreement: ‘It hardly seems plausible that simply adding to Stevenson’s story that the two parties hate each other should explain their not disagreeing after all; indeed it seems perverse.’ (Impassioned Belief, 170-71). I find this case difficult to imagine without knowing why the parties have to eat together, but it is not obvious to me that the parties disagree. As our responses to the blackmail cases suggest, hostility can undermine rather than intensify disagreement; it can turn potential disagreeers into mere antagonists.

86 The first two lines of this example are from Justin Khoo and Joshua Knobe, ‘Moral Disagreement and Moral Semantics’, Nous (‘Early View’ version) (2016): 4.
to them at all. Perhaps—though the idea seems far-fetched—this is a merely linguistic disagreement, which simply consists in the utterance of imperative sentences with what we take to be mutually unsatisfiable contents, or the performance of conflicting speech acts. Sally and Cody might not be in a state of disagreement at all; they might merely be engaged in the activity of disagreement.  

Since the disagreement between Sally and Cody is robust, since it might be non-cognitive, and since the Entrusted Concerns Hypothesis lays down necessary conditions for robust non-cognitive disagreement, the case might be a counterexample to that hypothesis. A counterexample of this sort would not trouble me too much, for, given my metaethical ambitions, I am most concerned to identify sufficient conditions for robust non-cognitive disagreement. But if there were a kind of robust non-cognitive disagreement whose sufficient conditions differed from those stated by the Entrusted Concerns Hypothesis, I would certainly wish to explore it, for it might reveal new and better options for a non-cognitivist moral psychology.

In what follows I shall argue that we do not have good reason to think that intuitions of robust disagreement elicited by cases in which different speakers utter conflicting imperatives (henceforth ‘prima facie imperative disagreements’) are, or are explained by, intuitions about a kind of non-cognitive disagreement that is not comprehended by the Entrusted Concerns Hypothesis. Intuitions of robust disagreement elicited by prima facie imperative disagreement only seem to exist where the parties’ utterances are plausibly explained by their having disagreeing thoughts of a (by now) familiar kind—conflicting descriptive beliefs, conflicting normative beliefs, or conflicting voiceable preferences. Admittedly my argument is far from conclusive because I only have the space to consider a limited range of cases. Part of my argument relies on a promissory note, namely, that the Entrusted Concerns Hypothesis can explain our intuitions about third-party non-cognitive disagreements—disagreements between two people about the choices of a third party. I make good on this promise in Chapter 2.

The first thing I want to establish is that we might be able to explain our intuitions about Imperative Disagreement by hypothesizing that we assume that the speakers (i) share entrusted concern relationships, and (ii) are making demands of each other qua

---

87 However, if we took Sally and Cody to be merely engaged in the activity of disagreement, we might perhaps find it infelicitous to say that the two disagree, rather than that they were disagreeing. On the state/activity distinction see Herman Cappelen and John Hawthorne, *Relativism and Monadic Truth* (Oxford: Oxford University Press, 2009), 60-61.
parties to those relationships. On this view, our natural interpretation of the case is very similar to that which I suggested we give to Restaurant. We assume that the speakers share two entrusted concern relationships, each giving one party a voice in the other’s decisions as to whether or not he will consent to go where the first party wants to go, and hence drink her preferred kind of drink. The parties make their demands by uttering imperatives instead of explicitly voicing preferences, but it is their opposing preferences that put them in disagreement. This explanation is at least apparently available. So we may now ask, can it be ruled out? Or if it can’t, is a better explanation nonetheless available?

One way to rule out an Entrusted Concerns Hypothesis-style explanation would be to present a version of the case in which the parties could not plausibly be thought to share the required trust relationships, but which nonetheless elicited an intuition of robust disagreement. But this is unpromising as a method for undermining the Entrusted Concerns Hypothesis’s explanation for our judgements about Imperative Disagreement, and, more generally, for undermining the hypothesis’s explanation for our judgements about similar cases in which two people are enjoining each other to do different things together. This is because, as a general matter, the very fact that two people are enjoining each other to do different things together is evidence that they share some sort of understanding, some joint project, or some amicable relationship, in the context of which it would not be surprising if each expected the other to be somewhat responsive to her wishes.

There are exceptions to this generalization—cases in which people have reason to jointly coordinate their actions despite apparently lacking any connection that could be construed as an entrusted concern relationship. I will discuss three cases which fit this description, and in which parties are in prima facie imperative disagreement. I consider whether they are counterexamples to the Entrusted Concerns Hypothesis.

First, there are cases like the blackmail case that I considered in 1.2.4, in which wholly antagonistic parties have a motive to do something together. Indeed Restaurant (Blackmail I) is, I think, a prima facie imperative disagreement of just this sort. However, we are not inclined to think that the situation of Cody and Sally is of this sort; we would need to be given some new information about their context before it would even occur to us that their relationship was antagonistic. If we did interpret Drinks in this way, I presume that we would feel hesitant about supposing the parties to be in disagreement, as we are when we are presented with Restaurant (Blackmail I).
Second, suppose that a man has fallen into the sea, and sharks are closing in on him. Two perfect strangers see this, and have the following exchange.

_Rescue_

\[ \begin{align*}
A: & \quad \text{Let’s row out in that boat and rescue him.} \\
B: & \quad \text{No, let’s look for a rope or a lifebuoy and throw it to him.}
\end{align*} \]

This case is not at all puzzling, and the parties seem to disagree, even though we have reason to doubt that they share any special entrusted concern relationships. But now it is hard to rule out a different explanation which, like the Entrusted Concerns Hypothesis, does not require us to posit a distinctive kind of non-cognitive disagreement. To wit, it is hard to rule out the hypothesis that we intuit disagreement merely because we take the parties’ conflicting imperatives to be explained by disagreeing _beliefs_—specifically, judgements about which course of action would achieve the greatest expected preservation of life and limb.

A third exception: five trapped people will be crushed by a speeding trolley if two bystanders don’t cooperate to push a very large man into its path, killing the large man but saving the five. The bystanders, who are perfect strangers, have the following exchange:

_Trolley_

\[ \begin{align*}
A: & \quad \text{Let’s shove that big guy in front of the trolley before it kills them.} \\
B: & \quad \text{No, let’s not push him to his death.}
\end{align*} \]

Again this sounds natural, and sounds like a disagreement, notwithstanding the presumable lack of entrusted concern relationships between the speakers. But our intuition of disagreement here is plausibly explained by an assumption that the parties have conflicting moral judgements. Strangers seem to be willing to cooperate in pursuit of moral ends without having to form entrusted concern relationships (this, of course, is part of the explanation for the naturalness of the previous case too). This provides us with an explanation for A and B’s ‘mutual enjoinings’ that does not (obviously) attribute any entrusted concern relationships to them, but it also suggests an explanation for their disagreement that does not make it (obviously) non-cognitive, viz., the parties are in moral disagreement.

Prima facie imperative disagreements like _Drinks_, in which two people enjoin each other to do different things together, seem quite reliably to elicit intuitions of robust disagreement. But I am unable to think of one in which we cannot naturally
attribute a descriptive disagreement, normative disagreement, or disagreement in voiceable preferences to the speakers. Antagonistic cases like Restaurant (Blackmail I) are an exception: the parties in such cases do not seem to be in one of these familiar sorts of disagreements, but neither do they clearly disagree.

What about other kinds of imperative disagreement though? If we can find prima facie imperative disagreements of another kind which fail to elicit intuitions of robust disagreement when the disagreement cannot be explained as an intuition about a disagreement in normative judgements, descriptive beliefs, or voiceable preferences, this would cast serious doubt on the idea that prima facie imperative disagreements ever amount to non-cognitive disagreements of a distinctive kind. There are no natural-seeming prima facie imperative disagreements which consist simply in one person resisting the demand of another (as John does, on my stipulative sense of ‘demand’, in Football (Mother and Son)). A person does not resist another’s command by giving a contrary command to herself. And I take it that an imperative disagreement requires the utterance of two or more mutually unsatisfiable imperatives. The only other kind of prima facie imperative disagreement that I am aware of is one in which two people give mutually unsatisfiable commands to a third party, i.e. prima facie third-party imperative disagreements.

Disagreements of this sort that fail to elicit intuitions of robust disagreement are easy to come by. Suppose that rival hostesses Mrs. A and Mrs. B encounter Mr. Toff at the same time, and that the following discourse ensues:

Rival Hostesses (Imperatives)

Mrs A: [To Mr. Toff] Come to my party!
Mrs B: [To Mr. Toff] No, come to my party!

As in the earlier rival hostesses case, this seems to be a case of mere rivalry without disagreement. There is not plausibly any disagreement in moral judgements, descriptive beliefs, or voiceable preferences between Mrs. A and Mrs. B. And in the absence of such a disagreement they do not seem to disagree at all.

It might be thought that Mrs. A and Mrs. B do not seem to disagree because, as a general matter, prima facie third-party imperative disagreements do not elicit intuitions of disagreement, or at least not unless they indicate a moral disagreement or a disagreement in descriptive belief. If so, our intuitions about Hostesses (Imperatives) do not support the view that our intuitions about Drinks must be explained by our tacit
attribution of a clash of voiceable preferences to the speakers. But I don’t think this is right. Imagine that a young woman, Penny, is deciding whether to marry the well-connected Mr. Toff, or Mr. Lloyd, a man from a family that has shared a long alliance and friendship with Penny’s family. A marriage to Toff will improve the social standing of Penny’s family, but disappoint the Lloyds and weaken somewhat the ties between the families. Penny’s mother wants her family to rise in society; Penny’s aunt cares most about honouring and cementing their ties to the Lloyds. They address Penny thus:

\[\text{Marriage (Imperatives)}\]

\[\begin{align*}
\text{Mother:} & \quad [\text{To Penny}] \text{Marry Mr. Toff!} \\
\text{Aunt:} & \quad [\text{To Penny}] \text{No, marry Mr. Lloyd!}
\end{align*}\]

Penny’s mother and aunt do seem to disagree here. And this seems to be explained by the relationships between the parties, and their reasons for their injunctions. They are making demands in the service of ends that Penny and her mother and aunt presumably all have in common, even if they care about them to differing degrees. These features of the case support the thesis that the mother and aunt have a third-party disagreement in voiceable preferences, according to the theories concerning third-party non-cognitive disagreement and networks of trust relationships that I develop in Chapters 2 and 3.\(^88\) (I discuss this case specifically in Appendix D.)

It might be conceded that, absent normative or descriptive disagreement, intuitions of robust disagreement elicited by prima facie imperative disagreements only occur where my conditions for disagreement in voiceable preferences are met. But why think that these intuitions are explained by a clash of preferences in the right social conditions rather than an opposition of imperatives in the right social conditions? Perhaps all apparent robust disagreements in preference are fundamentally disagreements in imperative. Reports of wishes often have the illocutionary force of commands—perhaps voicing preferences in the relevant circumstances is just a polite or mild way of giving orders.

The main problem with this idea is that intuitions of disagreement elicited by cases like Football (Mother and Son) cannot be interpreted as disagreements in

\(^88\) There is, I concede, a kind of normative disagreement between Penny’s aunt and mother here: they disagree about what Penny should do. (As I argue in 2.1, all robust third-party non-cognitive disagreements are or involve normative disagreements.) But I contend that this is a special kind of normative disagreement that only members of Penny’s family can participate in, and that is best explained by the postulation of special familial entrusted concern relationships among those involved.
imperative because, as I have already observed, one party to the disagreement fails to utter any imperative, or make any preference report that could be construed as an imperative. John merely says that he prefers to play football. What kind of command might this be? If we hypothesize that mutually unsatisfiable voiceable preferences (absent yielding etc.) suffice for robust non-cognitive disagreement, we can seemingly explain all intuitive cases of robust non-cognitive disagreement; whereas if we hypothesize that mutually unsatisfiable imperatives in certain social conditions suffice for robust non-cognitive disagreement, we cannot. Explaining all robust non-cognitive disagreement as disagreement in voiceable preferences is economical and gets the facts right.\textsuperscript{89}

\textsuperscript{89} I also find it intuitively very plausible that parties to robust non-cognitive disagreements have disagreeing \textit{thoughts}—that they are in a psychological \textit{state} of disagreement; and that the hypothesis that robust non-cognitive disagreement is always disagreement in voiceable preferences, and never merely imperative disagreement, should be preferred because it explains this. But I won’t defend this intuition.
Third-Party Non-cognitive Disagreement

Two people S and T may be in moral disagreement with each other, not just about what S or T ought to do, but also about what a third party U ought to do. They can have moral disagreements about U’s actions, ones that do not boil down to mere disagreements about what they themselves should try and get U to do. In my terminology, they can have third-party disagreements. My non-cognitivist theory of moral psychology ought to predict this; it thus requires that people should be capable of having (robust) third-party non-cognitive disagreements.

Happily, such disagreements do seem to be possible. I have already argued that the intuition of disagreement elicited by Marriage (Imperatives) is explained by our assumption of an underlying disagreement in preferences. For a clearer example, consider a variation on Stevenson’s football case in which John’s father makes an appearance.

Football (Parents I)

Mother: John’s thinking about quitting the football team.
Father: But playing for St. Matthew’s is a family tradition!
Mother: I’d rather he quit. Football is dangerous. He could get injured.
Father: Well I’d rather he didn’t quit.

John’s parents seem to disagree. But while this appearance is encouraging, it at best shows that (absent yielding) moralists with opposing preferences might disagree, whereas my theory needs to predict that they would. I need to identify at least sufficient conditions for third-party non-cognitive disagreement, and these must be conditions such that I can credibly postulate that they obtain in moral contexts.

My aim in this chapter is to give a plausible account of the conditions for third-party non-cognitive disagreement. Explaining why, given this account, we expect these
conditions to be satisfied in intuitive non-moral cases of third-party-disagreement will turn out to be a very large task, and I will not address it until Chapter 3. In the present chapter I will proceed as follows. In Section 2.1, I will say a bit more about the phenomenon I aim to give an account of, to distinguish it from other apparently similar phenomena. In Section 2.2 I present my theory of third-party disagreement.

2.1 Identifying Third-Party Disagreements

A possible external symptom of third-party non-cognitive disagreement is that the parties will have exchanges whose form is some variation on: ‘I’d prefer that so-and-so did such-and-such’/ ‘I’d rather so-and-so did not do such-and-such’. However, exchanges of this sort, even when they elicit intuitions of disagreement, can also occur in the absence of true third-party non-cognitive disagreement. Suppose that John’s parents have the following exchange:

**Football (Parents II)**

*Mother:* John’s thinking about quitting the football team.
*Father:* That means he’ll be around the house a lot more after school doesn’t it? I won’t be able to relax.
*Mother:* I’d rather he quit. It means I wouldn’t have to keep driving him home after training.
*Father:* Well I’d rather he didn’t quit.
*Mother:* We disagree then.

This seems to be a disagreement, and it seems to concern what a third party is going to do, yet it does not seem to be a case of third-party non-cognitive disagreement. This is because the grounds the parties give for their preferences have nothing to do with John’s concerns or interests; they are all about what would suit the parties themselves. The parties appear to be saying things they would be less likely to say while John was around. Parents can typically influence their children’s choices, and I am inclined to think that the intended upshot of the conversation is either that at least one of the parties would discourage John from quitting (if John’s father has his way), or that both would forbear from so discouraging him (if his mother has her way). And whatever the discouragement involved, it would probably not involve John’s father voicing the same preferences that John’s father voices here, with the same grounds, in John’s presence. If this is right, there is a sense in which what the parties are really disagreeing about is how they should influence John, rather than about John’s actions.
By contrast, the parties in *Football (Parents I)* give the same sorts of grounds for their preferences that they might give if they were talking to John. As in *Football (Parents II)*, it is plausibly an intended upshot of the conversation that the parties should agree to discourage or forbear from discouraging John from quitting. But intuitively, they are not disagreeing about whether to discourage John from quitting—or not merely. They are disagreeing about John’s action.

The following observation seems to support the view that *Football (Parents I)* is, and *Football (Parents II)* is not, a third-party disagreement. It is very natural to say that, in *Football (Parents I)*, John’s parents disagree about whether John should quit the football team. This is not true of John’s parents in *Football (Parents II)*. To disagree about what John should do is unambiguously to have a third-party disagreement. Of course, we might wonder whether it is really to have a non-cognitive disagreement. The ‘should’ that John’s parents disagree about, so to speak, is normative, not like the hypothetical ‘should’ in ‘X prefers that Y should φ’. John’s parents are having a normative disagreement. Cognitivists about the normative will want to say that John’s parents disagree in (cognitive) normative belief, and may or may not think that they also disagree in preference. I contend that John’s parents have a disagreement in preference, and that this disagreement just is a normative disagreement.

It will perhaps not come as a surprise that I have no quick way at my disposal to decide between these interpretations. I might suggest that the fact that the parties are explicitly voicing preferences rather than beliefs warrants a presumption that their disagreement is non-cognitive. And I might demand to know how a cognitivist would explain the apparent dependence of John’s parents’ disagreement on the relationship between them and John. Why is it that, if a well-meaning stranger were to chime in with his preference that John should not quit the team, on the same grounds that John’s father offers, he would not succeed in disagreeing with John’s mother? Can’t he make normative judgements about whether John should quit? Or is it that his preference-reports don’t conventionally indicate such judgements? If so, why not? The cognitivist has some explaining to do here, but I have no reason to think that the explanation would prove difficult.

So I am reduced to postulating that third-party disagreements such as that illustrated in *Football (Parents I)*, which are normative and which have the appearance of disagreements in preference, are in fact non-cognitive disagreements. The proposition that they are non-cognitive is a claim of my metaethical theory, rather than a datum that
supports the theory. It is nonetheless, I think, quite a plausible proposition, considered in isolation from its broader philosophical implications.

Let me return to my explanation of why the appearance that *Football (Parents I)* is and *Football (Parents II)* is not a disagreement about what John should do supports the view that only the former is a third-party disagreement. Take the set of disagreements that, like those depicted in *Football (Parents I)* and *Football (Parents II)*, have the verbal form of third-party non-cognitive disagreements. I contend that, within this set, those disagreements that are genuine third-party non-cognitive disagreements are those that are about what the third party should do (or ought to do etc.). Intuitively, these are the only disagreements in the set that are really about the actions of third parties. We have no words with which to say precisely what third-party non-cognitive disagreements are *about* if we cannot use normative terms—no way of completing the sentence, ‘S and U disagree about whether T _____ φ or ψ’, which implies that S and T are in third-party disagreement, without using ‘ought to’, ‘should’, ‘must’, and so on. If two people disagree about what a third party is to do, they are seemingly disagreeing about what to make the third party do. This is not a normative disagreement (or needn’t be), but neither is it a third-party disagreement. I cannot think of a way to say what actions the disagreement between John’s parents in *Football (Parents II)* is about that makes it a disagreement about John’s actions rather than those of his parents, because the disagreement cannot be characterized as one about what John should do.

Before I move on, another apparent distinguishing feature of third-party disagreements bears mentioning, although it is not a surprising one given the distinguishing feature I have already proposed, and doesn’t permit us to state the subject matter of such disagreements in non-normative terms. Observe that it is plausible that John’s mother in *Football (Parents I)* approves, and that his father disapproves, of John’s quitting the football team. Their preference reports seem to communicate approval and disapproval. By contrast, the preferences reported in *Football (Parents II)* do not seem to express approval or disapproval. I suspect that we will almost always find that the parties to third-party non-cognitive disagreements plausibly disapprove of

---

90 Cases of non-third-party non-cognitive disagreement can properly be described as disagreements about what the agent is to do, without any implication that it is not really that agent’s actions that are being disagreed about. This is because, although the parties to such a disagreement are disagreeing about what, between them, they will make the agent do, this very disagreement is a disagreement about the agent’s actions, since the agent is himself one of those parties.
their dispreferred actions, and tend to approve of their preferred options.91 (The sketch for an analysis of approval and disapproval presented in 5.3.2 suggests that parties to third-party disagreements might sometimes disapprove of their preferred as well as their dispreferred options.)

2.2 A Theory of Third-Party Disagreement

2.2.1 Waiving

In Section 2.2 I will develop and defend my theory of third-party non-cognitive disagreement. According to the theory, all such disagreements involve non-third-party disagreements about conventional actions that I will call acts of ‘waiving’. These are actions that can be performed by entrusting parties in entrusted concern relationships. In this subsection I will say what waiving is, and why conventions permitting it would plausibly exist. Before I do so it will be helpful to introduce some more terms with which to categorize the desires and preferences that affect the choices of parties to entrusted concern relationships. I will call the desires an entrusted party has to avoid straining and violating his partner’s trust, his trust-based desires. I will assume that a person’s trust-based desires qua entrusted party in a certain entrusted concern relationship are never among his weighted desires in that relationship.92

---

91 Crucial to the prediction that third-party disagreements involve clashing attitudes of approval and disapproval is my claim that third-party disagreements only occur among parties to trust networks whose members have shared ends, where these shared ends are suitably reflected in their weighted desire-types. Following Philippa Foot, I maintain (5.3.2) that one of the marks of approval (in contrast to the attitude of ‘being for’, for instance) is that it is based on concerns that the approved is presumed to share with the approver. See Foot, Virtues and Vices, 200-201.

92 If his trust-based desires were among his weighted desires, this would undermine the point of his having discretion-licensing preferences, which is to serve as a counterweight to his obligations, and make his membership in the trust relationship less burdensome. It would also make possible situations in which an entrusted party would be unable to relieve the strain he had put on his partner’s trust, no matter how often he acquiesced to her preferences. Recall Stevenson’s original football case. Suppose that John’s trust-based desires were among his weighted desires, and that he has strained his mother’s trust to the point that, if he resists her preferences again, he will violate her trust. If John is genuinely committed to living up to his mother’s trust, this means that he will have a trust-based desire to accommodate her preferences in his next decision that is strong enough to give him an actual, all-things-considered preference to do as she prefers, no matter what options his other desires might incline him to choose. Thus, since this trust-based desire is weighted, once he knows his mother’s accommodation-obliging preference, he would never have a discretion-licensing preference to do otherwise than his mother prefers. This means that he will never be able to yield to her preferences, for an entrusted party can only yield to a preference when he has a discretion-licensing preference to resist it. And since, normally, entrusted parties can only relieve strain by yielding, there is no way in which John could ever hope to return to a state in which he would sometimes enjoy
Because a party’s voiceable preferences may be (and, in the case of entrusted parties, must be) based on a proper subset of her actual desires, her voiceable preferences among a set of options may differ from her ordinary all-things-considered preferences among those options. An entrusted party’s voiceable preferences may also differ from those preferences he would have, taking account of all his desires except his trust-based desires. I shall call preferences of this last sort *pre-trust preferences*, because they are his preferences ‘before’ he takes considerations of trust into account.

If, out of two mutually exclusive, jointly exhaustive options, parties to an entrusted concern relationship both have voiceable preferences for the same option, they do not disagree, and the entrusted party must choose that option or violate his partner’s trust. This may sometimes be an unwelcome situation for both parties. The option for which the entrusting party has an all-things-considered preference and for which the entrusted party has a pre-trust preference may differ from the option they both voiceably prefer. To permit an escape from this situation, there would, I suggest, be a conventional action that the entrusting party could perform that would relieve the entrusted party from his obligation to accommodate her preference, and hence from his obligation to act in accordance with his discretion-licencing preference. I will not call this action a species of yielding, because I want to preserve the idea that yielding is acquiescing to a voiceable preference. Yielding by an entrusting party is acquiescing to her partner’s acting on his voiceable preference, but the sort of conventional action I am now talking about would typically involve allowing the entrusted party to act against his voiceable preference. So I will instead refer to it as *waiving*, in a particular decision, an entitlement to accommodation (or, more loosely, ‘waiving a demand’).

I will assume that waiving can only be performed by an entrusting party when her voiceable preference and that of her partner are for the same option, or when her partner lacks a discretion-licencing preference. A party may lack a voiceable preference concerning a particular decision if she has indeterminate utilities among the options; or the expected utilities of the options, in terms of weighted desire satisfaction, are determinate but equal for her (i.e. she is in a ‘Buridan’s ass’ situation); or she has no weighted desires relevant to the decision. An entrusted party who lacks a voiceable discretion to resist his mother’s preferences, unless he renegotiates his trust relationship with his mother.
preference must act on his partner’s voiceable preference or violate her trust, unless she waives. I shall revise my account of waiving in 3.3.3, and again in Appendix E.

2.2.2 The Four Relationships Hypothesis

I should warn the reader that the material from this subsection until the end of Part I will be complicated and difficult, certainly the most difficult part of this dissertation. However, the material is vital to my case for the existence of the moral trust network, and my explanation of third-party disagreement, among other phenomena.

I call my theory of the conditions for third-party non-cognitive disagreements the Four Relationships Hypothesis, because it posits that such disagreements can only occur where four entrusted concern relationships exist among the disagreeing parties and the agent whose actions are in dispute. This theory is based on the idea that part of what makes John’s parents’ clashing preferences in Football (Parents I) a disagreement about what John should do is that, as I observed above, these are preferences that they might also have voiced in John’s presence, on the same grounds, to urge him to act accordingly. If two people voice disagreeing non-counterfactual judgements about what someone else should do, then they voice judgements that they think the third party would recognize as applying to him, and as being of potential practical significance to him. They do not take themselves to be disagreeing about something of no interest to the agent whose actions they are talking about, like scientists discussing the behaviour of a beetle. Any account of third-party non-cognitive disagreement must save this appearance. The Four Relationships Hypothesis posits a network of entrusted concern relationships that explains why a disagreeing party voices a preference concerning what a third party should do, why this preference is of interest to the other disagreeing party, and why it is of practical significance to the agent.

According to the Four Relationships Hypothesis, a state of third-party non-cognitive disagreement differs qualitatively from a state of non-third-party non-cognitive disagreement, in being a ‘higher-order’ state. If we hold trust relationships fixed, third-party non-cognitive disagreement supervenes on non-third-party non-cognitive disagreement. Suppose that S prefers that U should φ, that T prefers that U should not-φ, and that these preferences put S and T in a state of third-party non-cognitive disagreement. For this to be the case, U must be a common entrusted party to S and T—that is, U must be the entrusted party in two entrusted concern relationships, one with S, another with T—and two entrusted concern relationships must exist.
between S and T, one with S as the entrusting party, the other with T as the entrusting party. The third-party non-cognitive disagreement between S and T is realized by a non-third-party non-cognitive disagreement in one or the other of these relationships between S and T, depending on whether S or T disagrees with U about whether or not U is to \( \phi \).

The following five criteria are intended to be necessary and jointly sufficient conditions for third-party non-cognitive disagreement. As usual, ‘prefer’, ‘preference’, etc. mean ‘voiceably prefer’, ‘voiceable preference’, etc. Since I shall have nothing to say about cognitive disagreements here, I won’t bother to specify that the disagreements I am discussing are non-cognitive. As I have mentioned, four entrusted concern relationships must exist for third-party disagreement to be possible, each with a different entrusted party/entrusting party pair. In what follows, \( \Delta_{(X,Y,Z,\phi)} \) stands for ‘the disagreement between X and Y about whether Z should \( \phi \)’. I shall use the following notation to refer to the entrusted concern relationships posited by the Four Relationships Hypothesis to explain a third-party disagreement \( \Delta_{(X,Y,Z,\phi)} \): where V and W are different people; V is X, Y, or Z; and W is X, Y, or Z; ‘\( E(V,W,\Delta_{(X,Y,Z,\phi)} \)’ stands for the ‘the entrusted concern relationship in which V is the entrusting party and W the entrusted party, on whose existence disagreement \( \Delta_{(X,Y,Z,\phi)} \) depends’. I assume throughout that preferences are rational.

**Condition One.** The third party whose options a third-party disagreement concerns must be a common entrusted party to the disagreeing parties. Weighted desire-types are the same for this common entrusted party in each relationship.

That is: necessarily, if disagreement \( \Delta_{(S,T,U,\phi)} \) exists, then S is the entrusting party in an entrusted concern relationship (namely, \( E(S,U,\Delta_{(S,T,U,\phi)} \)) with U entitling S to a voice in U’s choice to \( \phi \) or not-\( \phi \), and T is the entrusting party in an entrusted concern relationship (\( E(T,U,\Delta_{(S,T,U,\phi)} \)) with U entitling T to a voice in U’s choice to \( \phi \) or not-\( \phi \); and the weighted desire-types for U in these two relationships are identical.

**Condition Two.** There is a two-way entrusted concern relationship between the parties to a third-party disagreement entitling each to a voice in the other’s waiving choices in their relationships with the third party. Each disagreeing party has the same weighted desires in one relationship as she has in the other.

Why, and under what circumstances, would one party X to a third-party disagreement want the other party Y to waive qua partner to the third party Z? In general, X will want Y to waive if Y prefers, and X disprefers, the option that Z prefers.
Suppose that $X$ prefers that $Z$ should not-φ, and that both $Y$ and $Z$ prefer that $Z$ should φ. If $Y$ hasn’t waived, then $Z$ cannot yield to $X$’s preference without violating $Y$’s trust, because this would mean defying $Y$’s preference despite lacking a discretion-licencing preference to do so. If $Y$ has waived, then $Z$ can yield to $X$’s preference in defiance of $Y$’s without violating $Y$’s trust. The probability that $Z$ not-φs will thus be at least slightly higher. This is why $X$ will generally want $Y$ to waive if $X$ disagrees and $Y$ ‘agrees’ in preference with $Z$.

A more precise statement of Condition Two: necessarily, if $\Delta(S,T,U,\phi)$ exists, there is an entrusted concern relationship $(E(S,T,\Delta(S,T,U,\phi)))$ between $S$ and $T$ entitling $S$ to a voice in $T$’s choice to waive or not waive qua $U$’s partner (when $T$ has an entitlement to accommodation that $T$ can waive, i.e., when $T$ and $U$ both have preferences for the same option in $E(T,U,\Delta(S,T,U,\phi))$ concerning $U$’s choice to φ or not-φ); and there is a second entrusted concern relationship $(E(T,S,\Delta(S,T,U,\phi)))$ between them, of opposite direction, entitling $T$ to a voice in $S$’s choice to waive or not waive qua $U$’s partner (when $S$ has an entitlement to accommodation that $S$ can waive, i.e., when $S$ and $U$ both have preferences for the same option in $E(S,U,\Delta(S,T,U,\phi))$ concerning $U$’s choice to φ or not-φ); and the desire-types that are weighted for $S$ in $E(S,T,\Delta(S,T,U,\phi))$ and $E(T,S,\Delta(S,T,U,\phi))$ are identical; and the desire-types that are weighted for $T$ in $E(S,T,\Delta(S,T,U,\phi))$ and $E(T,S,\Delta(S,T,U,\phi))$ are identical.

**Condition Three.** A third-party disagreement between two people about the actions of a third is realized by pairs of non-third-party disagreements among those three people. At least one of these non-third-party disagreements must be between the parties to the third-party disagreement.

More precisely: necessarily, if $\Delta(S,T,U,\phi)$ exists, then either

(i) $U$ has a discretion-licencing preference to φ or not-φ in $E(S,U,\Delta(S,T,U,\phi))$ and $E(T,U,\Delta(S,T,U,\phi))$, and either

(a) $S$ disagrees with $U$ qua party to $E(S,U,\Delta(S,T,U,\phi))$ about whether $U$ is to φ, and $S$ and $T$ disagree qua parties to $E(T,S,\Delta(S,T,U,\phi))$ about whether $T$ is to waive in $E(T,U,\Delta(S,T,U,\phi))$; or

(b) $T$ disagrees with $U$ qua party to $E(T,U,\Delta(S,T,U,\phi))$ about whether $U$ is to φ, and $S$ and $T$ disagree qua parties to $E(S,T,\Delta(S,T,U,\phi))$ about whether $S$ is to waive in $E(S,U,\Delta(S,T,U,\phi))$; or

(ii) $U$ lacks a discretion-licencing preference to φ or not-φ in $E(S,U,\Delta(S,T,U,\phi))$ and $E(T,U,\Delta(S,T,U,\phi))$, and $S$ and $T$ disagree qua party to $E(S,T,\Delta(S,T,U,\phi))$ about whether $T$ is to waive in $E(T,U,\Delta(S,T,U,\phi))$, and $S$ and $T$ disagree qua party to $E(T,S,\Delta(S,T,U,\phi))$ about whether $S$ is to waive in $E(S,U,\Delta(S,T,U,\phi))$. 

78
**Condition Four.** Weighted desires for the disagreeing parties qua parties to their two shared entrusted concern relationships are identical to their weighted desires qua partners to the common entrusted party. That is: necessarily, if S and T have a third-party disagreement Δ about U’s choice to φ or not-φ, then the desire-types that are weighted for S in \(E(S,T,Δ_{S(T,U,φ)})\) and \(E(T,S,Δ_{S(T,U,φ)})\) and \(E(S,U,Δ_{S(T,U,φ)})\) are identical; and the desire-types that are weighted for T in \(E(S,T,Δ_{S(T,U,φ)})\) and \(E(T,S,Δ_{S(T,U,φ)})\) and \(E(T,U,Δ_{S(T,U,φ)})\) are identical (see Figure 1).

![Figure 1](attachment:image1.png)

Figure 1. Trust relationships that must exist between, and weighted desire-types of, S, T, and U, if S and T are in non-cognitive disagreement about whether or not U should φ. Arrows between individuals indicate entrusted concern relationships, and point from entrusting party to entrusted party. W, X, and Y, are sets of desire-types. Their placement at the beginning or end of an arrow indicates that they are weighted for the party at that end of the arrow in the relationship signified by that arrow.

**Condition Five.** Where X and Y are parties to a third-party disagreement about what Z should do, X must have preferences concerning choices to waive or not waive qua Z’s partner that are ‘in harmony’ with X’s preferences regarding Z’s actions. It must not be possible for X to (rationally) both prefer, qua partner to Z, that Z should φ, and, qua partner to Y, to lack an accommodation-obliging preference that Y should waive if both Y and Z prefer that Z should not-φ; nor to lack a discretion-licensing preference to not-waive if both X and Z prefer that Z should φ.

Note that the satisfaction of Condition Five is not guaranteed by that of Condition Four, because the considerations that explain X’s preference that Z should φ may not exhaust the considerations bearing on X’s preferences concerning whether or not to waive qua Z’s partner if Z also prefers φ, or on whether or not Y should waive qua Z’s partner if Y and Z prefer that Z should not-φ. For instance, even if we suppose that X’s waiving qua Z’s partner when (Z prefers to φ and Y prefers that Z not-φ) always raises the probability that Z not-φs, X might (for all Condition Four says) prefer to waive (despite her preference that Z should φ) because it would make X seem interestingly
unpredictable; this is a consideration that does not bear upon X’s preference that Z should φ. In light of these observations, it will not be clear what it would take to satisfy Condition Five. I will address this question in 3.3.3.

Here is a more careful statement of the condition. Necessarily, if \( \Delta(S,T,U,\varphi) \) exists, then the terms of \( E(S,T,\Delta(S,T,U,\varphi)) \), \( E(T,S,\Delta(S,T,U,\varphi)) \), \( E(U,S,\Delta(S,T,U,\varphi)) \), and \( E(T,U,\Delta(S,T,U,\varphi)) \) are such that, necessarily (holding fixed everything except S’s and T’s beliefs and desires), if U prefers to φ qua party to \( E(S,U,\Delta(S,T,U,\varphi)) \) and \( E(T,U,\Delta(S,T,U,\varphi)) \), and S and T have voiceable preferences qua parties to \( E(S,T,\Delta(S,T,U,\varphi)) \) or \( E(T,S,\Delta(S,T,U,\varphi)) \), then either

(i) (a) S qua party to \( E(S,U,\Delta(S,T,U,\varphi)) \) prefers that U φ, and
(b) T qua party to \( E(T,U,\Delta(S,T,U,\varphi)) \) prefers that U not-φ, and
(c) S prefers qua party to \( E(T,S,\Delta(S,T,U,\varphi)) \) that S not waive in \( E(S,U,\Delta(S,T,U,\varphi)) \), and
(d) T prefers qua party to \( E(T,S,\Delta(S,T,U,\varphi)) \) that S waive in \( E(S,U,\Delta(S,T,U,\varphi)) \); or

(ii) (a) S qua party to \( E(S,U,\Delta(S,T,U,\varphi)) \) prefers that U not-φ, and
(b) T qua party to \( E(T,U,\Delta(S,T,U,\varphi)) \) prefers that U φ, and
(c) S prefers qua party to \( E(T,S,\Delta(S,T,U,\varphi)) \) that T waive in \( E(T,U,\Delta(S,T,U,\varphi)) \), and
(d) T prefers qua party to \( E(S,T,\Delta(S,T,U,\varphi)) \) that T not waive in \( E(T,U,\Delta(S,T,U,\varphi)) \); or

(iii) (a) S qua party to \( E(S,U,\Delta(S,T,U,\varphi)) \) prefers that U φ, and
(b) T qua party to \( E(T,U,\Delta(S,T,U,\varphi)) \) prefers that U φ; or
(c) S prefers qua party to \( E(T,S,\Delta(S,T,U,\varphi)) \) that S not waive in \( E(S,U,\Delta(S,T,U,\varphi)) \), and
(d) T prefers qua party to \( E(T,S,\Delta(S,T,U,\varphi)) \) that T not waive in \( E(T,U,\Delta(S,T,U,\varphi)) \); or

(iv) (a) S qua party to \( E(S,U,\Delta(S,T,U,\varphi)) \) prefers that U not-φ, and
(b) T qua party to \( E(T,U,\Delta(S,T,U,\varphi)) \) prefers that U not-φ.

(Lines (i) and (ii) describe states in which S and T have a third-party disagreement; lines (iii) and (iv) describe states in which they do not.) This concludes my statement of the conditions for third-party disagreement.

Condition Three ensures that the parties to a third-party disagreement actually disagree with each other according to the Entrusted Concerns Hypothesis. Because of this condition, cases of third-party disagreements ought to elicit intuitions that the putative disagreers actually disagree. But the main challenge I face in accounting for the phenomenon of third-party disagreement is to explain why the things that count as third-party disagreements according to the Four Relationships Hypothesis can properly be said to be disagreements about the actions of the third party. To meet this challenge I must complete two tasks: I must show that the things that the Four Relationships hypothesis counts as third-party disagreements are intuitively disagreements about the
third party’s actions. And I must show that, if the parties to one of these things were to express their disagreeing preferences, they would plausibly voice preferences that were about the entrusted parties’ actions, rather than about the waiving choices of one or the other party.

One obvious point I can make with regard to the first task is that, according to the Four Relationships Hypothesis, if S and T have a third-party disagreement about whether U should φ, they necessarily have conflicting voiceable preferences about whether or not at least one of S and T should waive qua U’s partner. Condition Three requires this; Condition Two guarantees that they share the trust relationships that entitle them to voice these preferences. So some of their preferences have something to do with bringing it about that U is more or less likely to φ. I don’t think that this suffices to make their disagreement one about whether U should φ, but it intuitively brings it into a closer resemblance to such a disagreement than it would have had, had their preferences been about something completely unrelated, or about (e.g.) whether or not they themselves should φ in U’s situation. Condition Two explains how S and T could, at least in some sense, be at odds with each other about U’s actions; Condition Three requires that they would.

The network of trust relationships required by Conditions One and Two, and the ‘harmony’ among the preferences of the parties in each of the relationships required particularly by Conditions Four and Five, allows me to address the second task. These explain why a third-party disagreement can properly be described as a clash among preferences concerning what the third party should do,93 and hence why the parties would express their disagreement by voicing preferences concerning what the third party should do.

Suppose that S has a third-party disagreement Δ with T about whether U should φ. In accordance with Condition Three, S and T are parties to one or two (if U lacks a preference) non-third-party disagreements about whether or not one of them is to waive qua U’s partner. (Call the one- or two-member set of these disagreements D.) But common sense tells us that if two people disagree in preference about the actions of a third party, the preferences they voice will be preferences about those actions, not about the waiving choices of one of the parties to the disagreement. Why, if they are in disagreement(s) D, should S and T voice preferences about whether U should φ, rather

93 This is a non-normative ‘should’.
than about whether S or T should waive qua U’s partner? Let’s start by explaining why S and T can communicate to each other the preferences that put them in D by voicing preferences concerning whether U should φ, then consider why they would.

The ‘harmony’ required by Conditions Four and Five among S’s and T’s preferences in $E(S,T,\Delta(S,T,U,\phi))$, $E(T,S,\Delta(S,T,U,\phi))$, $E(S,U,\Delta(S,T,U,\phi))$, and $E(T,U,\Delta(S,T,U,\phi))$ means that if S voices a preference that U should φ, T can infer that S has the preference that opposes T’s in D, whatever D is about. If U prefers to not-φ, then D is a disagreement about whether T should waive (qua U’s partner, of course), and T can infer that S prefers that T should waive. If U prefers to φ, then D is a disagreement about whether S should waive, and T can infer that S prefers not to waive. If U lacks a preference between φ-ing and not-φ-ing, then D is a set of two disagreements, one about whether T should waive, the other about whether S should waive, and T can infer that S prefers that T should waive in the first, and that S should not waive in the second. So S and T can communicate to each other the preferences that put them in D by voicing preferences concerning whether U should φ—the preferences that put them in third-party disagreement.

Why would they voice these preferences? D is a disagreement about what S and/or T is to do; surely it would be weirdly indirect to voice preferences concerning what U should do instead. But in a way, the specific actions that D is about are not what put S and T at odds with each other. Suppose the following. U prefers to not-φ; D is thus a disagreement about whether T should waive. U subsequently changes her mind, and comes to prefer to φ. Now S and T are no longer in D; they no longer disagree about whether T should waive. But they are still in non-third-party disagreement—this time about whether S should waive. And exactly the same desires, and the same considerations concerning U’s actions, put them in disagreement. So voicing preferences concerning what U should do is not really indirect—rather, it avoids complications about who is to waive qua U’s partner, and places the focus on what robustly puts S and T at odds. Thus for most purposes, S’s and T’s disagreement is most aptly described as a disagreement about what U should do.\footnote{I shall be forced to revise this part of my explanation of third-party non-cognitive disagreement. See Appendix E.}

The robustness of third-party disagreements, according to the Four Relationships Hypothesis, with respect to changes in the third party’s preferences is thus an important
part of what makes them third party disagreements. It is to predict this robustness that the hypothesis requires there to be four entrusted concern relationships connecting the two disagreers and the third party, even though a non-third-party disagreement only ever exists in two of these relationships at a time. Why wouldn’t three relationships suffice instead of four? Suppose that each of S and T is an entrusting party in an entrusted concern relationship with U (call these E(S,U) and E(T,U)), that T is an entrusted party in an entrusted concern relationship with S (call this E(S,T)) which gives S a voice in T’s waiving choices qua U’s partner. These are the only entrusted concern relationships connecting the three; importantly, T does not have a voice in S’s waiving choices. For simplicity’s sake, suppose that the weighted desire-types for S, T, and U in all of these relationships are the same, and that there is a ‘harmony’ constraint, similar to that imposed by Condition Five, that prevents S from preferring that T should not waive when S disagrees with U in E(S,U) and T ‘agrees’ with U in E(T,U), etc.

Now suppose the following: in E(S,U) and E(T,U), U prefers to φ; in E(S,U) S prefers that U should not-φ; in E(T,U) T prefers that U should φ; and in E(S,T) S prefers that T should waive qua U’s partner (as ‘harmony’ requires). No one has yielded or waived. S and T disagree. What is wrong with the hypothesis that this is a third-party disagreement? The problem is that this disagreement is specifically about whether T should waive, for it persists only as long U’s preferences are such that T’s waiving would increase the likelihood that U would not-φ—only as long as U has a preference in E(T,U) and E(S,U) that ‘agrees’ with T’s and disagrees with S’s. If U’s preference changes, then S and T are no longer at odds, for T’s waiving would not affect U’s choices, and, though S’s waiving might, T lacks a voice in S’s waiving choices.

I said above that the Four Relationships Hypothesis aims to capture a necessary or at least typical feature of normative disagreements about the actions of third parties, namely, that if people voice disagreeing judgements about what someone else should do (or ought to do, etc.), then they voice judgements that they think the third party would recognize as applying to him. It will perhaps be obvious by now how the hypothesis purports to do this. I have already explained why the parties S and T to a third-party disagreement Δ would voice preferences concerning what the third party U should do rather than what about what waiving choices they themselves should make. The ‘harmony’ required by Conditions Four and Five among S’s and T’s preferences in E(S,T,Δ(S,T,U,φ)), E(T,S,Δ(S,T,U,φ)), E(S,U,Δ(S,T,U,φ)), and E(T,U,Δ(S,T,U,φ)) means that these preferences are also S’s and T’s accommodation-obliging preferences in their
relationships with U. Thus the preferences two people will voice qua parties to a third-party disagreement are the same preferences that the third party is obliged to accommodate if he is to live up to their trust, unless he has discretion-licensing preferences of his own. Thus he would recognize the preferences the disagreers voice as applying to him.

In summary, the Four Relationships Hypothesis predicts that we would have an intuition of robust non-cognitive disagreement when we are presented with a case of third-party disagreement, because it requires that there should be at least one robust non-third-party non-cognitive disagreement between the parties to a third-party disagreement. It explains why S and T would, in at least a vague sense, be at odds with each other concerning U’s φ-ing, by requiring that they be in non-third-party disagreement about whether one or both of them should waive their entitlement to demand that U should φ or not-φ. It explains why, despite the fact that this non-third-party disagreement concerns whether one or both of S and T should waive their entitlement to a voice in U’s φ-ing, it is nonetheless most apt, for normal purposes, to describe them as being in disagreement about whether U should φ, and as so disagreeing in virtue of holding conflicting preferences concerning whether U should φ. And it explains why S and T would assume that the preferences that put them in third-party disagreement are ones that would have practical significance for U himself. Because of this, and assuming that we tacitly infer that Conditions One to Five hold in any intuitive case of non-cognitive third-party disagreement, I contend that these features suffice to explain the phenomenon of non-cognitive third-party disagreement.

Of course this is a big assumption, and as long as it remains just an assumption I can’t claim to have adequately explained the phenomenon. In the next chapter I will try to explain, for at least a small range of apparent third-party disagreements, why it is reasonable to assume that the four requisite relationships among the disagreers and the agent would obtain. Because my aims are metaethical, and because I am postulating that moral judgements are possible in virtue of moralists’ membership of what I call ‘dual-order fully connected networks’, I will focus my attention on third-party disagreement in networks of entrusted concern relationships of this sort. I address the question of why we plausibly assume that the four appropriate relationships exist in Football (Parents I) and Marriage (Imperatives) in Appendices C and D, but my explanations assume knowledge of Chapter 3.
Networks of Trust

In Part II I will argue that moralists are members of a certain kind of network of entrusted concern relationships. Each member is an entrusted party and an entrusting party in an entrusted concern relationship with each other member. The weighted desire-types for all parties in all these relationships are identical. The relationships are different from those we considered in Chapter 1, for if a member of this network wants to change another’s behaviour, she could rarely if ever hope to do so by simply demanding, qua entrusting party, that he do as she prefers; usually, the best strategy available to her will be to change the reasoning behind his discretion-licencing preferences. Partly because of the homogeneity of weighted desire-types among moralists, if she succeeds in changing his discretion-licencing preferences, she will change his accommodation-obliging preferences too, and so will change the demands he makes of others. Moralists engage with the reasoning behind the preferences of other moralists even when they are not contemplating performing any actions. They will reason with each other about the actions of others, and have third-party non-cognitive disagreements.

These features of the moral trust network are supposed to explain, or partly explain, basic facts about our moral practices: moralists reason with each other about what they should do instead of merely making demands of each other; they argue together about what others should do; and they demand that all other moralists perform (or forbear from performing) the same behaviours, and strive to exhibit these behaviours themselves. My aim in this chapter will be to explain why a network of trust relationships with these particular features of my postulated moral trust network would exist and how it would work, and also to show, where I can, that I am not postulating anything too unfamiliar.
I shall thus have three concerns in this chapter: to identify a type of trust network that permits its members to have third-party disagreements according to the Four Relationships Hypothesis defended in Chapter 2, and has the other features of the moral trust network that I have mentioned; to identify the conditions under which we should expect such a trust network to exist; and, along the way, to give examples of trust networks with some or all of these features where possible.

I shall start by considering a simple kind of trust network, whose members consist of a single entrusted party and multiple entrusting parties. I call this a ‘common entrusted party network’. The relationships among the members of a common entrusted party network can make competing demands of the common entrusted party, potentially making the network unstable. In Subsection 3.1.2 I identify two ways in which this problem can be avoided. The application of one of these solutions to the problem of instability produces common entrusted party networks that have important features in common with the moral trust network. I identify an example of such a network in 1.3. In 3.1.4 I describe the features of a salient version of this sort of network, which I call a ‘uniform weighted desires network’. I explain why we should expect such networks to exist. Trust networks more elaborate than uniform weighted desires networks are needed to allow entrusting parties to disagree about the actions of a common entrusted party. In 3.2.3 I identify a kind of trust network consisting of many ‘overlapping’ uniform weighted desires networks which does permit third-party non-cognitive disagreement, and has all of the features of the moral trust network that I alluded to above. I describe the circumstances in which we should expect to find these networks. Throughout these sections I largely ignore the problem of how Condition Five of the Four Relationships Hypothesis could be met—of how ‘harmony’ could be achieved. I finally address this problem in Section 3.3.

Before proceeding, I must say something about my methods and assumptions in this chapter, and the breadth of my brush. Throughout this chapter I shall be making empirical claims about the trust relationships that tend to exist among human beings in various circumstances. I shall appeal to two types of considerations to support these claims: facts about what would serve the interests of members of the hypothesized relationships; and intuitions—mainly about disagreement, about who has voices in whose decisions, and about what sorts of grounds certain people can appropriately cite in support of their voiced preferences. These intuitions support my empirical hypotheses only if certain other propositions are credible. These include the theories about
disagreement and the social bases for having a voice a decision that I defended in Chapters 1 and 2. They also include the proposition that our intuitions about disagreement and about voiceable preferences are accurate. That is, such intuitions, concerning a certain case, are well correlated with the facts about who would actually disagree with whom, and who would actually have voiceable preferences in whose decisions, if the case were real. I shall assume that our intuitions are accurate in this way in what follows. This is a plausible assumption: it would be odd if our intuitions about who has voices in whose decisions were very inaccurate, given the importance that these sorts of social facts presumably have for our welfare and reproductive success.

Not all of the assumptions I have made in this chapter are plausible. Some are most likely false, and are moreover inconsistent with claims I have committed to elsewhere. The material in this chapter is very complicated—it is the most difficult part of the thesis. I have chosen to make simplifying assumptions to avoid complicating it further. The most important of these will only become apparent late in this dissertation, when I revise my account of the obligations and entitlements of entrusted parties. These revisions are relevant to the success of the arguments I make in this chapter, but I have chosen to postpone them and ignore them here for simplicity’s sake. In Section 3.3, I have also chosen to ignore certain complications arising from the fact that entrusted parties in large networks will be ignorant of many of their partners’ preferences.

3.1 Conditions for Stability in Common Entrusted Party Networks

3.1.1 The Main Source of Conflict

A person may, of course, be an entrusted party in more than one entrusted concern relationship at the same time (my Four Relationships Hypothesis presupposed this). He may have decisions to make in which several entrusting parties have voices. For instance, a young woman who wants to marry a certain man may find that this preference puts her in disagreement with various relatives who would rather she married someone else. Where multiple entrusting parties have a common partner, and both have

95 In 8.2.1 I argue, very roughly, that entrusted parties are only obliged to do what their evidence indicates would best satisfy their partners’ weighted desires, and that they only have discretion to resist doing this if it is to perform the actions that their evidence indicates would best satisfy their desires.
voices in the same decisions of that common partner, they and he are members of what I will call a *common entrusted party network*.

Common entrusted party networks will be unstable unless the terms of the trust relationships are constrained in certain ways. An example will illustrate the main source of instability. Suppose that A1 and A2 are entrusting parties who share the same partner B, and have voices, qua B’s partner, in the same set of decisions. A1 prefers that B should φ; A2 prefers that B should not-φ; neither entrusting party is willing to yield. Presumably, different desire-types might be weighted for B in B’s relationships qua partner to A1 and A2. Accordingly, B may have discretion-licencing preferences favouring different options in these relationships. Here are the possible combinations of B’s preferences between φ and not-φ:

<table>
<thead>
<tr>
<th>Qua A1’s partner</th>
<th>Qua A2’s partner</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) B prefers φ</td>
<td>B prefers φ</td>
</tr>
<tr>
<td>(2) B prefers φ</td>
<td>B prefers not-φ</td>
</tr>
<tr>
<td>(3) B prefers not-φ</td>
<td>B prefers φ</td>
</tr>
<tr>
<td>(4) B prefers not-φ</td>
<td>B prefers not-φ</td>
</tr>
</tbody>
</table>

Let’s consider the consequences of these possible combinations.

In situation (1), B must either φ or violate A1’s trust. He has a discretion-licencing preference to φ in both relationships. He could yield to A2’s preference that he not-φ, but if he does so, he will defy A1’s preference despite lacking a discretion-licencing preference to not-φ, which violates the terms of his trust relationship with A1. In (2) B cannot avoid violating either A1’s trust or A2’s. His preferences in each trust relationship agree with those of his partner, but he cannot satisfy both of these preferences, because they are for incompatible actions. If he φs he defies both A2’s preference, and his own discretion-licencing preference qua A2’s partner. If he not-φs he defies both A1’s preference, and his own discretion-licencing preference qua A1’s partner. There are no problems in situation (3). B can yield to either A1 or A2 without defying her discretion-licencing preferences in the other relationship. Situation (4) mirrors situation (1); B must either not-φ or violate A1’s trust. If he yields to A2’s preference and not-φs, he defies both A1’s preference and his own discretion-licencing preference.

These outcomes are summarized in this expanded version of the table above:
Qua A1’s partner Qua A2’s partner Outcome
(1) B prefers φ B prefers φ B must not yield (he must φ)
(2) B prefers φ B prefers not-φ B must violate someone’s trust
(3) B prefers not-φ B prefers φ B must yield to either A1 or A2
(4) B prefers not-φ B prefers not-φ B must not yield (he must not-φ)

In situation (2) one of B’s trust relationships is imperilled no matter what B does. If (2)-type situations are possible in a common entrusted party network, the network is unstable. It does not make sense for both entrusting parties to trust the common entrusted party (assuming they know of each other’s relationship with him), nor for the entrusted party to commit to accommodate both entrusting parties’ preferences. So we should expect actual common entrusted party networks, if they exist, to rule out (2)-type situations somehow.

3.1.2 Two Ways to Rule Out (2)-Type Situations

As far as I can see, there are two options for ruling out (2)-type situations. First, (2) can only occur if B can have different discretion-licencing preferences qua A1’s partner and qua A2’s partner. This possibility would be ruled out if the terms of B’s relationships with both A1 and A2 counted exactly the same desire-types as weighted for B. Under this arrangement, if A1 and A2 had opposing preferences, the trust network would not necessarily be threatened, for B could act on his discretion-licencing preference in both relationships without violating anyone’s trust. B could not yield to either party however. Or rather, he could never yield to one party in such a situation without violating the other’s trust. Every choice situation he faced qua entrusted party, when A1’s and A2’s preferences were for incompatible options, would resemble possibility (1) or (4). The acceptability of this sort of arrangement might depend on how many entrusting parties were sharing the same partner. The more entrusting parties there were, the less likely it would be that they would ever all have preferences for the same option, and the less likely it would be that the common entrusted party would ever yield to any of them. Call this the uniform discretion-licencing desires solution, and call a common entrusted party network employing this solution a uniform discretion-licencing desires network. (‘Discretion-licencing desire’ is not part of my established terminology; it refers to a weighted desire of an entrusted party. I shall not use it except in the labels introduced above.)

According to fiduciary non-cognitivism, (2)-type situations are prevented from arising in the moral trust network by means of this solution. Since I do not want my
postulated moral trust network to seem exotic in this regard, I shall try to provide evidence that non-moral trust networks employing this solution also exist. One of my aims in this subsection will be to establish what would constitute evidence of this sort.

Consider now the second way of preventing (2)-type situations. B presumably has trust-desires not to violate A1 and A2’s trust. These desires could be weighted for B; they could be, as I shall say, weighted trust-based desires. In 1.2.2 I argued that, as a general matter, entrusted parties’ trust-based desires are not weighted. But by this I meant that an entrusted party’s trust-based desires in a particular entrusted concern relationship would not be weighted for him in that relationship. The possibility that I am now broaching is that the entrusted party in a common entrusted party network might have trust-based desires qua party to one relationship in the network that are weighted for him in other relationships in the network.

Actually, talk of ‘weighted trust-based desires’ is a bit broad for our purposes. Let me introduce a distinction among trust-based desires. An entrusted party’s accommodative trust-based desires are, I shall say, those of his trust-based desires that incline him to accommodate his partner’s preferences. An entrusted party’s discretional trust-based desires are those of his trust-based desires that incline him not to exceed his discretion (i.e. not to do anything beyond what his rational discretion-licencing preferences licence him to do). Thus his accommodative desires incline him to act on his partner’s voiceable preferences, and his discretion-based desires incline him to act on his own voiceable preferences (rather than defy both them and his partner’s preference). Note that an entrusted party does not exceed his discretion when he yields to his partner’s voiceable preference in defiance of his own, so his discretion-based desires will never incline him to resist yielding. Accommodative and discretion-based desires need not be basic, and may be derived from substantially the same more-basic desires, such as the entrusted party’s desire to preserve his trust relationship. (Thus the idea that B might have weighted trust-based desires contradicts my hypothesis that weighted desires must be basic—more on this shortly.)

According to the second solution to the instability problem, B would have weighted discretion-based desires. More precisely, B’s desire not to violate A1’s trust by

---

96 ‘Discretional desires’ are not to be confused with ‘discretion-licencing desires’. I am avoiding the more familiar word ‘discretionary’ to avoid the suggestions, which it seems to encourage, that these desires are (i) desires that an entrusted party possesses because he has discretion to possess them, (ii) desires that he is permitted act upon because he enjoys discretion, or (iii) desires he has to exploit his discretion.
exceeding his discretion might be weighted for him qua A2’s partner, and his desire not to violate A2’s trust by exceeding his discretion might be weighted for him qua A1’s partner. If this were the case, then (2)-type situations would be impossible. Suppose that a *prima facie* (2)-type situation eventuated—that is, a situation in which A1 preferred φ, A2 preferred not-φ, and B, *bracketing his weighted discretional desires*, preferred φ qua A1’s partner and not-φ qua A2’s partner. ‘Unbracketing’ B’s weighted discretional desires, we find that B would have a weighted desire qua A1’s partner not to violate A2’s trust, and a weighted desire qua A1’s partner not to violate A1’s trust. These would suffice to give B discretion-licencing preferences in each relationship to do what the entrusted party in the other relationship wanted him to do. As a consequence, B would find himself in an unproblematic (3)-type situation rather than a genuine (2)-type situation. Call this the *weighted discretional desires solution*.

Note that my claim that this solution makes (2)-type situations impossible was based on the assumption that B’s weighted discretional desires would be strong enough to turn any prima facie (2)-type situation into a (3)-type situation. This assumption will admittedly not always hold. But the weighted discretional desires solution resolves the problems raised by (2)-type situations even if it does not rule those situations out. (2)-type situations initially seemed problematic because they forced B to violate trust no matter how much he desired to avoid it. B would suffer the (e.g. reputational) penalties for untrustworthiness, no matter how trustworthy he was, and this would make becoming a common partner to A1 and A2 an ill bargain to B. By contrast, the possibility that B might be ‘forced’ to violate someone’s trust because he didn’t care enough about living up to that trust seems far less problematic. If B suffers a loss of reputation through violating trust where this is the result of a lack of trust-based motivation, this loss of reputation does not seem arbitrary, for this lack of motivation is indeed evidence that B is not trustworthy (or, in less normatively loaded terms, that trusting B is risky). If B knew himself to be trustworthy, the possibility of his ending up in (2)-type situations would appear to him too remote to inhibit him from becoming a common partner to A1 and A2. So although the weighted discretional desires solution does not eliminate the possibility of (2)-type situations entirely, it does make this possibility less counterproductive.

Variations on the weighted discretional desires solution are possible. We might call the version of the solution described above the *symmetrical* weighted discretional desires solution, for it proposes that B’s desires to live up to each entrusting party’s trust
would be weighted for him qua each other entrusting party’s partner. Asymmetrical versions of the solution are possible too. Perhaps B’s desire not to violate A1’s trust would be weighted in his relationship with A2, but his desire not to violate A2’s trust would not be weighted in his relationship with A1. If so, B would always be obliged to yield to A1, never to A2, in prima facie (2)-type situations. This would turn prima facie (2)-type situations into (1)-type situations, rather than (3)-type situations, as the symmetrical version of the solution does. The asymmetrical version would make sense if B had an incentive to be more deferential to A1 than A2, e.g., if A1 were B’s parent, and A2 were a more distant relative. In groups where the number of entrusting parties with a common partner exceeds two, various mixtures of asymmetrical and symmetrical versions of the weighted discretional desires solution are possible.

What are the drawbacks of the weighted discretional desires solution? Although I intend to find fault with this solution in order to establish a place for the uniform discretion-licencing desires solution, there is one ‘problem’ with the weighted discretional desires solution that I wish to ignore. This is that, whereas weighted desires must, as I argued in 1.2.2, be basic, trust-based desires may be (and I think typically are) derived from other more fundamental desires. I don’t want to reject the weighted discretional desires solution on this ground because my argument that weighted desires must be basic is too weak to lean on. Moreover, even if this argument were correct, there might be a way to capture the spirit of the weighted discretional desires solution by identifying the basic desires from which discretional desires are derived, and hypothesizing that these would be weighted.97

A genuine problem of the symmetrical versions of the weighted discretional desires solution is that they may seem to give too much licence to entrusted parties. Suppose that John’s mother and father both have voices in John’s decision as to whether he plays football, and that they have opposing voiceable preferences: John’s father prefers that he play; John’s mother prefers that he not play. Suppose further that this is a prima facie (2)-type situation: bracketing weighted discretional desires, John has a discretion-licencing preference qua his father’s partner to do as his father prefers and a discretion-licencing preference qua his mother’s partner to do as she prefers. John also

97 These basic desires might need to specified in such a way as to ensure that they would not influence B’s preference otherwise than by the support that they gave to his discretional desires. Restrictions on admissible grounds, such as I discuss in Section 3, might be of use in this regard.
has a non-weighted desire to play football because he wants to emulate the popular kids. Because of this desire, he has a pre-trust preference to play football (he would not have had such a preference otherwise). Finally, the symmetrical weighted discretionary desires solution is in effect: John has weighted desires to live up to his mother’s trust qua his father’s partner, and to live up to his father’s trust qua his mother’s partner. This means that John is free to choose which of his parents he will yield to (as in a (3)-type case). This choice is unconstrained by trust-based considerations, so he can, and presumably will, choose to play football so that he can emulate the popular kids. It might seem undesirable to John’s parents that he should be free to make his choice on these sorts of grounds, whenever they had opposing preferences concerning what he should do. The uniform discretion-licencing desires solution might seem preferable here, because it would mean that John would at least have to choose on the basis of his weighted desires.

This problem of the weighted discretionary desires solution does not strike me as very serious in contexts in which only a few entrusting parties share a common partner (and have voices in the same decisions). It becomes more serious as the number of entrusting parties increases. For as it increases, the number of differing sets of desire-types that are weighted for B in his different relationships increases, and the probability that different entrusting parties in the network will have mutually unsatisfiable voiceable preferences increases. And so the likelihood that B will find himself in a prima facie (2)-type situation increases. That is, there will be an increasing likelihood that, bracketing B’s weighted discretionary desires, there will be some A1 and some A2 among B’s partners, and some φ, such that A1 prefers that B should φ, B prefers qua A1’s partner that he should φ, A2 prefers that B should not-φ, and B prefers qua A2’s partner that he should not-φ. If this ever occurs, then B will be free to choose to φ or not-φ on the basis of his pre-trust desires, and his weighted discretionary desires will licence him to do so in all of his relationships. The end result is that B’s actions will only occasionally be constrained by his discretion-licencing preferences or by the demands that particular entrusting parties make of him; he will very often be able simply to act on the promptings of his pre-trust desires, without fear of violating trust. Thus the trust relationships between B and his partners will influence his behaviour only to a limited degree, and only in a haphazard way.

I may seem to have exaggerated the problems of the weighted discretionary desires solution a bit here, by ignoring the fact that, when B exploits the licence that this
solution gives him in prima facie (2)-type situations, he *strains the trust* of at least one of his partners. The implementation of the weighted discrecional desires solution gives B a discretion-licencing preference to resist whichever demands he wants to resist in prima facie (2)-type situations, but, as ever (at least in the kinds of entrusted concern relationships we have considered so far), resisting entrusting parties’ accommodation-obliging preferences involves straining their trust. So B is not freed from trust-related pressures altogether in these situations.

This would be a stronger point in favour of the weighted discrecional desires solution were the influence of ‘strain’ on an entrusted party’s situation not itself rather problematic in contexts in which a large number of entrusting parties shared a partner. For such entrusting parties will frequently have preferences for the common entrusted party to choose incompatible options, and the entrusted party has no choice in these situations but to act on her discretion-licencing preferences qua partner to some subset of the entrusting parties, and so strain the trust of this subset. And, no matter how judiciously the entrusted party chooses, he is bound, over the course of a long enough series of decisions, to violate some entrusting party’s trust. This is counterproductive. Trust violations diminish one’s reputation, and if being a partner to all of these entrusting parties means being unable to avoid violating trust, occupying this role will be unattractive. The obvious way to overcome this problem is to greatly weaken the entrusted parties’ obligations to accommodate the preferences of (most of) the entrusting parties, so that he can defy them again and again, in conformity to his discretion-licencing preferences, without significantly straining their trust. But this move means that the entrusted party will once again have too much licence if the weighted discrecional desires solution is in effect.98

Giving the common entrusted party licence to indulge his pre-trust preferences to a large extent might not always be unsatisfactory. If the entrusting parties in a common entrusted party network lack the power to induce the entrusted party to accept stringent terms, they might content themselves with permissive terms, as long their trust

98 In 1.5 below I postulate that entrusted parties may have ‘conflict excuses’ for resisting preferences, permitting them to defy their partners’ preferences without straining their trust when those preferences are for incompatible options. This would eliminate the need for entrusted parties to have only very weak obligations to avoid straining their partners’ trust. But it would also eliminate, on most occasions, the strain-based constraint on an entrusted party’s behaviour, that we supposed might mitigate the ‘too much licence’ problem I have raised for the symmetrical version of the Weighted Discrecional Desires solution.
relationships give them a modicum of influence over the entrusted party. But there are better and worse ways to be permissive. And in a common entrusted party network that relied wholly on the symmetrical version of the weighted discretional desires solution to avoid the problems posed by (2)-type situations, the entrusted party would enjoy great latitude under circumstances that were quite fortuitous and unpredictable. If permissiveness was called for, it would be better to achieve it by diminishing the range of decisions in which the entrusting parties had voices, rather than by making the circumstances of his freedom random. So the symmetrical version of the weighted discretional desires solution seems wanting. In common entrusted party networks with many entrusting parties, we should expect the problem posed by (2)-type situations to be addressed by either the uniform discretion-licencing desires solution or an asymmetrical version of the weighted discretional desires solution.

I am not aware of any general problems facing asymmetrical versions of the weighted discretional desires solution. It may make sense in some common entrusted party networks, if the entrusting parties have different levels of power over the common entrusted party. In such contexts, an asymmetrical version of the weighted discretional desires solution would allow the entrusting parties a degree of influence over the entrusted party that approximately reflected the varying degrees of their underlying power over him.

I am now in a position to say what it would take to provide evidence that uniform discretion-licencing desires networks exist ‘in the wild’ rather than as a mere possibility. If each party in a network had equal underlying power over the common entrusted party, the asymmetrical weighted discretional desires solution would give entrusting parties unequal influence over him in an arbitrary way, and so would, I presume, be unsatisfactory. A symmetrical version of the solution would not have this problem, but would have the defect I discussed earlier: it would give the entrusted party too much freedom to do as he pleased, only occasionally constraining him to act on his discretion-licencing preferences or defer to his partners’ preferences. So I can provide evidence for the existence of common entrusted party networks exploiting the uniform discretion-licencing desires solution by giving a plausible example of a largish common entrusted party network all of whose entrusting parties have equal underlying power over the entrusted party.
3.1.3  *Uniform Discretion-Licencing Desires Networks in the Wild*

The following elaborate case is supposed to meet these conditions.

The Seymours are an old, wealthy, and prominent family in a certain rural locality. They traditionally organize and partly finance an annual spring fair involving maypole dancing. Pete Seymour, a crowd-pleasing man and a former professional sportsman, presides as the chair of the Spring Fair Planning Committee. Pete’s lack of concern for family tradition and his perceived commonness have earned him the resentment of many Seymours. When he announces his plan to remove the maypole dance from the fair programme and replace it with a cricket match, his continued role as chair of the of the Spring Fair Planning Committee becomes all but intolerable to the other Seymours. His great popularity in the area makes his removal from the Committee difficult to engineer, however. Bertie Seymour happens to acquire photographs of Pete behaving scandalously while drunk. Bertie could send the photos to the press. Doing so would make it easy to get Pete fired from the Committee, but would embarrass the Seymour family.

Bertie is a Seymour in more than name: he is a proud member of the family, hospitably disposed towards its members and friends, and ill-disposed toward those who threaten it. He is a financially independent adult. Things will go better for him if the other Seymours think well of him, but none of them enjoys any special power over him.

The Seymours, or at least a large subset of them, would plausibly enjoy voices in Bertie’s choice of whether or not to send the photographs to the press. If we imagine various Seymours approaching him, telling him they know about the photos, voicing preferences for him to do one thing or another, and citing as grounds for their preferences facts about the reputation and traditions of the Seymour family, their utterances will not seem presumptuous. If we imagine Bertie voicing contrary preferences in reply, citing similar facts, he will seem to disagree. This supports the hypothesis that a common entrusted party network would exists among the Seymours with Bertie as the common entrusted party if the case were real. (This line of support depends on the assumptions, discussed in the introduction to this chapter, that my theories about entrusted concern relationships and non-cognitive disagreement are correct and that our intuitions about who has voices in whose decisions are accurate.) My stipulation that none of the Seymours has a special degree of power over Bertie should discourage the idea that an asymmetrical version of the weighted discretional desires solution was in play. So it is reasonable to infer that, if Bertie’s case were real,
there would probably be a common entrusted party network among the Seymours centred on Bertie, and that (2)-type situations would be prevented in this network exclusively by means of the uniform discretion-licencing desires solution.

3.1.4 The Properties and Salience of Uniform Weighted Desires Networks

In this subsection I will describe a class of uniform discretion-licencing desires networks. I call the networks in this class uniform weighted desires networks, because their essential distinguishing feature is that the weighted desire-types of all members in all relationships are the same, not just the entrusted party’s ‘discretion-licencing’ desires. I will also argue that this class is salient among uniform discretion-licencing desires networks: a disproportionate range of these will be uniform weighted desires networks. This is because situations favouring the formation of uniform discretion-licencing desires networks very commonly also favour the formation of uniform weighted desires networks. As I shall argue, the common entrusted party network that we assume exists in Bertie’s case is a uniform weighted desires network, and I shall refer to this case extensively, both as an illustration of the properties of uniform weighted desires networks and in my argument for their salience. I am interested in uniform weighted desires networks because according to fiduciary non-cognitivism the moral trust network is a complex of uniform weighted desires networks. By giving evidence for the existence of non-moral uniform weighted desires networks, I show my postulation of the moral trust network to be less exotic. As usual, ‘prefers’ (etc.) means ‘voiceably prefers’ (etc.) in what follows, unless I indicate otherwise.

Before I discuss uniform weighted desires networks specifically, let me point out two (typical) properties of uniform discretion-licencing desires networks. The first is a necessary feature of such networks, the second is a property we should contingently expect them to have. Here is the first:

**UDLD1.** The entrusted party can never yield to any of his partners’ preferences without violating someone’s trust, unless each one of his partners either prefers the option that the entrusted party disprefers, or has waived her entitlement, or lacks a relevant voiceable preference.

Suppose that, in his relationships, a common entrusted party in a uniform discretion-licencing desires network, B, prefers to φ. (He must, of course, prefer the same option in all his relationships, since his weighted desires are the same in all of them.) One and only one of the following must be true: among B’s partners, there is

(i) some partner who (prefers that B not-φ and has not yielded), and
some partner who (prefers that $B \phi$ and has not waived); or

(ii) some partner who (prefers that $B \neg \phi$ and has not yielded), and
no partner who (prefers that $B \phi$ and has not waived); or

(iii) no partner who (prefers that $B \neg \phi$ and has not yielded), and
some partner who (prefers that $B \phi$ and has not waived); or

(iv) no partner who (prefers that $B \neg \phi$ and has not yielded), and
no partner who (prefers that $B \phi$ and has not waived).

I contend that $B$ can yield without exceeding his discretion only if either (ii) or (iv) is true. In (i) $B$ can yield but would violate the trust of at least one partner who has not waived. In (ii) $B$ can yield, and would not exceed his discretion in any of his relationships by doing so. In (iii) and (iv) $B$ cannot yield because there is no one for him to yield to. So $B$ can yield without exceeding his discretion only if (ii) is true. A fortiori he can yield without exceeding his discretion only if (ii) or (iv) is true. Thus, as I have claimed, the entrusted party in a uniform discretion-licencing desires network cannot yield without violating someone’s trust, unless each one of his partners either prefers the option that the entrusted party disprefers, or has waived her entitlement, or lacks a relevant voiceable preference.

What this means is that yielding by an entrusted party will be a relatively unusual thing in a uniform discretion-licencing desires network. And it will tend to be rarer the larger the network is, because situations in which there is no partner who both prefers the same option as the entrusted party and has not waived will be less likely in larger networks.

I’ll now turn to the second, contingent, property of uniform discretion-licencing desires networks:

UDLD2. The terms of the trust relationships in the network are such that the entrusted party cannot significantly strain his partners’ trust by failing to yield to their preferences, unless each one of his partners either prefers the option that he disprefers, or has waived her entitlement, or lacks a relevant voiceable preference.

I foreshadowed my claim that uniform discretion-licencing networks would have UDLD2 in 3.1.2 when I argued that the influence of ‘strain’ on entrusted parties in common entrusted party networks would be a source of instability. I explain in Appendix B why I have included the ‘unless…’ exception clause;\(^99\) we can ignore it for

\(^{99}\) In brief: I will postulate that the terms of the trust relationships among members of a uniform weighted desires network (a kind of uniform discretion-licencing desires network) may be such
present purposes. Let us consider why the network among the Seymours in Bertie’s case would probably have this property—why Bertie could not significantly strain trust by failing to yield.\(^{100}\) I shall assume in the remainder of this subsection the conclusion I argued for in the last: that the common entrusted party network centred on Bertie is a uniform discretion-licencing desires network.

It follows from this assumption that the network has UDLD1: Bertie cannot yield to his partners’ preferences without violating someone’s trust, unless those preferences are unanimously opposed to his, Bertie will rarely have the opportunity to yield. Most of the time he will just have to act in accordance with his discretion-licencing preferences, which will often involve resisting his partners’ preferences. If repeatedly resisting his partners’ preferences strains and eventually breaches their trust, he will ‘accumulate strain’ by consistently acting on his discretion-licencing preferences, and eventually breach some partner’s trust. This might occur even if he has never had the opportunity to relieve strain by yielding (without at the same time violating some party’s trust by defying both his own preferences and those of that party). As a general matter, the possibility of ending up in such a situation would make being the entrusted party in a large common entrusted party network unattractive, so it is reasonable to suppose that if a person occupied this role, the terms of his trust relationships would be such that he would not significantly strain trust by resisting his partners’ preferences (though he would, of course, violate trust by simultaneously defying his voiceable preferences and those of any partner).\(^{101}\)

Because uniform weighted desires networks are uniform discretion-licencing desires networks, uniform weighted desires networks will necessarily have UDLD1 and typically have UDLD2. I’ll now introduce the distinctive properties of uniform weighted desires networks, then argue for the ‘salience’ of these networks through a consideration of Bertie’s case. A network must have UWD1 if it is to count as a uniform

---

\(^{100}\) In what follows, when I talk about ‘Bertie’s relationships’, ‘Bertie’s partners’, etc., I have in mind only the relationships Bertie shares, and the partners he has, qua member of the common entrusted party network I have been discussing, and not any other e.g. more personal trust relationships he might be a party to.

\(^{101}\) My postulation of ‘conflict excuses’ in Appendix B will require a qualification of this generalization, but that postulation is not important to my arguments in this chapter.
weighted desires network; UWD2 is merely a property that we should contingently expect it to have given that it has UDLD2 and UWD1.

**UWD1.** The weighted desire-types for all the entrusting parties in the network are the same as the weighted desire-types for the entrusted party.

**UWD2.** Entrusting parties aspire to influence the entrusted party’s behaviour by moving him to change his discretion-licencing preferences rather than by inducing him to yield to their preferences.

I’ll defend simultaneously the propositions that UWD1 and UWD2 would be true of Bertie’s case, since the plausibility of each of these propositions depends at least to some extent on the plausibility of the other. An obvious upshot of my assumption that the network centred on Bertie is a uniform discretion-licencing desires network, and my observation that UDLD1 holds in such networks, is that entrusting parties in Bertie’s case are less likely to affect Bertie’s choices by getting him to yield the more of them there are, because unanimity of preferences is less probable in larger groups. This raises the question of why anyone would want to be one of a large number of entrusting parties in this network, and why such parties would bother to voice preferences.

Part of the attraction of being an entrusting party in Bertie’s network is that it gives one what I shall call ‘discursive entitlements’. Entrusting parties in a network of this sort are like all entrusting parties in that they are entitled to demand that their partners state what preferences they are acting upon, and their grounds for holding those preferences. Entrusting parties in general have this entitlement partly so that they can better detect abuses—so that they can better determine whether their partners’ voiced preferences are sincere, or whether they have just been voiced opportunistically to defend actions that serve the entrusted parties’ private interests. Knowing the history of an entrusted party’s voiced preferences, and his professed grounds for those preferences, an entrusting party can form a theory about his beliefs and desires in light of which to assess the likely sincerity of the preferences he voices subsequently. By this means she may better ensure that he is, to the required degree, acting to serve the interests that their trust relationship was formed to promote.

An entrusting party is also entitled to have her accommodation-obliging preferences heard and given consideration by the entrusted party. If she voices her preference and supplies grounds for it, the entrusted party must be able to explain why the reasoning behind her preference does not make it rational for him to hold a
preference of the same sort. The stronger the similarity between the weighted desire-types of the entrusting party and the entrusted party; and the greater the extent to which entrusting and entrusted parties, among the weighted desire-types that are common to both, may be assumed to have desires of these types that are similar in strength; the stronger the expectation will be that the reasoning behind the entrusting party’s preference should hold good for the entrusted party’s. Thus, where these similarities of type and strength are known to exist, the entrusting party’s ‘discursive entitlement’ to have her partner give consideration to her accommodation-obliging preferences becomes a means—albeit a highly fallible one—of influencing the entrusted party’s discretion-licensing preferences, and hence his actions, so as to bring them into line with the entrusting party’s accommodation-obliging preferences. For entrusting parties in common entrusted party network such as the one centred on Bertie, this is really the only ‘institutional’ means available (the ‘institution’ here being the common entrusted party network) of changing the entrusted party’s behaviour, save when he faces a decision concerning which all entrusting parties who have preferences and have not waived are unanimous in their preferences.

We come at last to my case for UWD1 and UWD2. That UWD2 is true of Bertie’s case is the point I’ve just made. Bertie’s partners won’t generally hope to get him to yield, because this option will not usually be available to him unless he is willing to violate someone’s trust. Rather they will aspire to change his discretion-licensing preferences by exercising their discursive entitlements. They will scrutinize his preferences, and challenge him to show that the reasoning behind their preferences is mistaken, or otherwise to explain why it does not give him the same preference.

We should expect UWD1 to be true in Bertie’s case because, as I have said, this sort of challenge will be more effective if potential differences between Bertie’s and the entrusting parties’ preferences arising from differences in their weighted desire-types are eliminated. The equality between Bertie’s and the entrusting parties’ weighted

---

102 It will turn out that it is because of this entitlement that my metaethical theory can partly accommodate Stevenson’s intuition that, ‘When you tell a man that he oughtn’t to steal, your object isn’t merely to let him know that people disapprove of stealing. You are attempting, rather, to get him to disapprove of it.’ (Charles L. Stevenson, ‘The Emotive Meaning of Moral Terms’, Mind 46, no. 181 (1937): 19); Stevenson himself could not well explain how moral utterances were supposed to change people’s non-cognitive attitudes (Dreier, ‘Relativism’, 103). On my theory, telling someone that he ought not to steal is voicing a preference qua entrusting party in a uniform weighted desires network that he should follow norms forbidding stealing. It is because of one’s entitlement that he should heed the reasoning behind one’s preference that one can hope to change his preference by so doing.
desire-types would be brought about by fitting the latter to the former rather than vice versa. Because Bertie’s weighted desire-types are the most important determinants of his behaviour qua entrusted party, we should expect them to be shaped by the underlying interests of and power relationships between Bertie and the entrusting parties. The entrusting parties’ accommodation-obliging preferences are comparatively unimportant, because of the obstacles to Bertie’s yielding to them, so they can be shaped to serve the more inconsequential end of facilitating the use of discursive entitlements to influence Bertie’s preferences, at no significant cost.

My argument for UWD1 and UWD2, applied to Bertie’s case, relies on the idea that being one of Bertie’s partners gives you a special entitlement to demand that Bertie give consideration to your preferences and their grounds, and explain his own preferences, so that the reasoning behind them could be examined. What exactly do I have in mind when I talk about the ‘reasoning’ behind the preferences of an entrusted or entrusting party? I mean at least the instrumental reasoning behind those preferences, which in this case involves the coherent proportioning of one’s voiceable preferences to one’s credences and weighted desires, approximately in accordance with the prescriptions of decision theory.

Factual reasoning about, for instance, the consequences of sending or not sending the photographs to the press, also seems to be relevant here. In general, an entrusting party seems to be entitled to demand that her partner explain the factual reasoning behind his preferences, and give consideration to the factual reasoning behind her preferences. Suppose that John from Stevenson’s football case tells his mother that he wants to play football, in defiance of her preferences, for some surprising reason—say, because he believes it will improve his schoolwork. John’s mother seems entitled to demand that he explain why he thinks that playing football will improve his schoolwork.

It is less clear to me that an outsider to their trust relationship lacks any similar entitlement. Recall Football (John and Well-Wisher). If John told Well-Wisher that playing football improved his schoolwork but, when asked, refused to explain why he believed this, he might seem to have violated a conversational norm obliging speakers to back up their assertions when certain conditions are met. Still, John’s mother’s entitlement to an explanation certainly seems stronger, and this ought to give her greater power to alter John’s discretion-licencing preferences through scrutinizing his reasoning than Well-Wisher enjoys. Moreover, while participants in a conversation might be
under some sort of obligation to back up their assertions, they aren’t necessarily obliged to make those assertions in the first place. Suppose that Well-Wisher has the same entitlement that John’s mother has to demand that John lay bare the reasoning behind the assertions he makes to explain his preferences, and hence the same capacity to change those preferences by pointing out flaws in his reasoning. Well-Wisher’s capacity to change John’s preference by scrutinizing John’s factual reasoning is nonetheless smaller than John’s mother’s, because Well-Wisher is not entitled to demand that John assert his relevant beliefs.

In this subsection I have argued that we have reason to think that UWD1 would be true of the trust network in Bertie’s case, and hence that the common entrusted party network centred on Bertie would be a uniform weighted desires network. My arguments for this conclusion have not depended on any peculiar features of Bertie’s case, but merely on two facts which ought to hold in any uniform discretion-licencing desires network: namely, that the common entrusted party would not be willing to put himself in a situation in which he could not help but over-strain trust (my case for UDLD2 applied to Bertie’s case), and that the entrusting parties would want their discursive entitlements to be more rather than less useful (my case for UWD2 applied to Bertie’s case). So we should expect uniform weighted desires networks to be disproportionately common among uniform discretion-licencing desires networks.

3.2 Trust Networks Permitting Third-Party Disagreement

In remainder of this chapter, my focus will be on establishing that there is a certain kind of uniform weighted desires network that plausibly exists, and that has the properties it must have for its members to have third-party non-cognitive disagreements about what other members should do, given the truth of the Four Relationships Hypothesis defended in Chapter 2. This kind of trust network is of interest because it subsumes the moral trust network I postulate in Part II. In the course of making this argument I hope to show that certain other properties of my proposed moral trust network would have precedents in smaller and humbler non-moral trust networks.

In this section I will first argue that, for all I’ve said so far, it is not clear that we should expect the conditions for third-party non-cognitive disagreement to typically obtain in uniform weighted desires networks. Second, I argue that there is nonetheless a certain subset of uniform weighted desires networks in which we should expect these conditions to obtain, and argue, of some uniform weighted desires networks whose
members seem capable of having third-party non-cognitive disagreements, that they belong to this subset. I will not complete this task in Section 3.2; an explanation of how the ‘harmony’ condition could be satisfied will be postponed until Section 3.3.

3.2.1 Why Might Conditions One to Five Be Met?

Entrusting parties in a uniform weighted desires network with opposing accommodation-obliging preferences concerning the actions of their common entrusted party are not necessarily in third-party disagreement, even if no one has yielded or waived, because they need not share the two-way entrusted concern relationships required by the Four Relationships Hypothesis. But they could share such relationships. And the possibility that they do must be one that we are quite willing to entertain, if the Four Relationships Hypothesis is correct, since we can easily imagine entrusting parties in Bertie’s case having such disagreements. Suppose that Bertie’s cousin Agnes Seymour wants the Seymour name to be more highly esteemed, and that his aunt Augusta Seymour cares a great deal about honouring the family’s traditions. Imagine the following conversation:

*Photographs*

*Bertie:* I could send these photos to *The Local Gazette*. Then we could get Pete ousted for certain.

*Cousin Agnes:* The scandal wouldn’t just embarrass Petey; it would embarrass us all.

*Aunt Augusta:* You’re right Agnes. But scrapping the maypole dance! It’s a local tradition—a Seymour tradition! I want Bertie to send the photos.

*Cousin Agnes:* Well I’d rather he didn’t. We disagree.

Cousin Agnes and Aunt Augusta are plausibly in non-cognitive disagreement about what Bertie should do.

I argue that there is an important subset of uniform weighted desires networks in which we should expect to find entrusting parties who are capable of having third-party disagreements about their common entrusted parties’ choices, but that in these cases we should also expect to find much ‘thicker’ networks of entrusted concern relationships than just those required by individual uniform weighted desires networks. I focus my attention on a salient subset of these, which I will call *fully connected networks*, to borrow a term from the field of network topology.103 (On my metaethical theory, the

moral trust network is an example of such a network.) Having explained why we should expect members of these networks to be capable of having third-party disagreements, I then explain why we expect the network among the Seymours to be a fully connected networks, and hence why *Photographs* elicits an intuition of disagreement.

I will call the facts about a group of individuals G and their environment that causally influence the arrangements of entrusted concern relationships we would find among G base conditions with respect to G. These will be facts about the desires of the individuals, especially the extent to which these desires would be satisfied by the same states of affairs (as I shall say, the extent to which they are common desires); the capacities of the different members of G to satisfy or frustrate these desires; and the distribution of ‘power’ within G. We can perhaps crudely think of a member’s ‘power’ as her relative capacity to satisfy or frustrate other members’ desires (i.e. any of their desires, not just common desires) without thereby frustrating her own. I’ll make quite a few simplifying assumptions about the features of a set of base conditions with respect to G in what follows. For instance I shall generally assume that members of G are able to communicate with one another, that they are not violently hostile to one another, and that their social setting is not of a very unusual sort. Note also that I will mainly be concerned with largish groups, comprising, say, ten or more individuals.

As a foil to my account of the base conditions for fully connected networks, let me describe what I take to be base conditions conducive to the development of a mere uniform weighted desires network, one whose members would not be able to disagree with each other about what the entrusted party should do. The conditions, among a group G, are as follows: (i) G has significant common desires W; (ii) there is only one agent S in G who can significantly affect the extent to which W are satisfied; (iii) non-B members of G collectively have some power over B (though not enough collectively to dictate her choices); and (iv) power (over each other and over B) is fairly evenly distributed among non-B members.

For all A such that A is a non-B member of G, we should expect A to entrust to B a concern to accommodate her W-type desires, when B’s preferences based on his W-type desires do not give him discretion to do otherwise, and we should expect B to commit to do so, because of A’s power over her. Because B would be the entrusted

‘fully connected networks’ are sometimes called ‘mesh networks’, though usage of the latter term seems to vary considerably.
party in an entrusted concern relationship of this sort with each other member of G, situations in which B has no choice but to violate some entrusting party’s trust are possible, unless the terms of B’s relationships are ‘designed’ to prevent such situations. Because the possibility of such situations is counterproductive, and because the alternative ways of avoiding them are flawed (for reasons that I discussed especially in 3.1.2), at least in common entrusted party networks whose entrusting-party members enjoy roughly equal power, we should expect the network of trust relationships centred on B to be a uniform weighted desires network. (Condition (iv) makes the existence of an entrusting party with disproportionate influence over B unlikely, and hence makes it unlikely that B would have discretional desires in any relationship that were weighted for him in his other relationships.)

We should not, I argue, expect B’s partners to form entrusted concern relationships amongst themselves giving them voices in each other’s waiving choices qua partners to B. This is not because it would not be in the interests of an entrusting party to have a voice in the other entrusting parties’ waiving choices. This would have advantages, though very slight ones. First, that party’s demands that other entrusting parties should waive might occasionally be successful, and thus, on rare occasions, she might be able to change a circumstance in which the common entrusted party could not yield to her preferences without violating other entrusting parties’ trust, into one in which the common entrusted party could yield to her preferences. By this means she could increase the probability that he would so yield. More importantly, having a voice in other entrusting parties’ waiving choices would give her discursive entitlements that they should lay bare the reasoning behind their preferences, and heed her own preferences and the reasoning behind them. This would allow her to bring their preferences into better alignment with her own by identifying flaws in their reasoning. It might also allow her to detect abuses on the part of entrusting parties whose voiced preferences opposed her own. By revealing that such parties had been deceitfully voicing mere pre-trust preferences as if they were accommodation-obliging, she could bring about their ostracism from the network, and so bring the possibility of the common entrusted party’s yielding to her preferences closer to reality.

The reason why we should not expect B’s partners to share two-way trust relationships of a sort that would make third-party disagreement possible is simply that they would have no real incentive to commit to accommodate each other’s preferences in their waiving choices. What does an entrusting party gain by granting other entrusting
parties the small advantages that I have cited, so that she might have the same advantages herself? Why would she give rival entrusting parties the means to slightly improve their chances of swaying the common entrusted party, so that she might have the same means at her disposal? There is little to be gained by such an arrangement.

Let’s now consider what I take to be the salient base conditions for a fully connected network among group G: (i) G has significant shared desires W, though there is considerable variation in the strengths of W-type desires among members of G; and (ii) all members of G can significantly affect the extent to which W are satisfied.

All members want to avoid frustrations of their W-type desires, but the ability of each member to frustrate these desires makes all vulnerable. They will thus want to form some sort of agreement whereby they will reciprocally forbear from frustrating W-type desires. The diversity in the strengths of W-type desires within G means that members will not all be willing simply to commit to follow a common comprehensive set of rules conformity to which would inhibit frustrations of their W-type desires. Rather, we would expect each member of G to entrust to each other member a concern to accommodate her W-type desires. Each member of G would be the common entrusted party in a uniform weighted desires network whose membership was G. They would form what I am calling a fully connected network.

Before I consider whether the members of G would have voices in one another’s waiving choices, let’s consider what form such entitlements would take if they did exist. One issue that must be addressed initially is whether they would enjoy these entitlements as parties to the network of trust relationships that I have just described, or as parties to an additional fully connected network concerned, presumably exclusively, with waiving choices. I shall proceed on the assumption that that an additional (‘second-order’) fully connected network is required. It might be possible to dispense with this second network, but showing this would require further argument, and perhaps a modification of the Entrusted Concerns Hypothesis that I would not otherwise need to commit to.

I will call the composite of a first-order and a second-order fully-connected network, of the sort that I am going to argue would exist among members of G, a dual-order fully connected network. The ‘first-order’ network within a dual-order network is, of course, the network that is concerned with concrete actions; the ‘second-order’ network is concerned with waiving choices. (By ‘concrete actions’ I mean actions other than the conventional actions such as yielding and waiving that can only be performed
by members of entrusted concern relationships.) This vocabulary may naturally be extended to particular entrusted concern relationships, and particular parties in those relationships. Thus a first-order entrusted concern relationship is one that is concerned with concrete actions; a second-order entrusted concern relationship is one concerned with waiving choices. An entrusting party in a first-order relationship is a first-order entrusting party, an entrusted party in a second-order relationship is a second-order entrusted party, and so on.

What desires would be weighted for parties to the second-order trust network? I suggest that the weighted desires for both entrusted and entrusting parties in this network, as in the first-order network, would be their W-type desires. W-type desires are, again, the desires that members of G hold in common. Exercises of entitlements by second-order entrusting parties would occasionally result in changes in a first-order entrusting party’s waiving choices, and the members of G would, I suppose, prefer that these waiving choices be made out of a concern to accommodate desires belonging to their common desire-types.

A much more important consideration supporting my claim that weighted desires in the second-order network would be W-type desires is the fact this would make ‘harmony’ among members’ preferences concerning both waiving choices and choices among concrete actions easier to achieve. ‘Harmony’ is desirable because it facilitates the scrutiny of preferences. It is generally useful for members of G if the preferences voiced by members of G in the first-order network are subjected to a good deal of scrutiny, both to determine that they are not means-end irrational, and to determine that they are genuinely voiceable preferences, and not ‘biased’ so as to serve members’ private interests. This is mainly because, if members act on instrumentally rational, ‘unbiased’, voiceable preferences, this will tend to serve the common desires of members of G better than acting on instrumentally irrational or biased preferences. Members will also be less willing to make the sacrifices the trust network requires of them to the extent that they believe other members are voicing and acting on ‘biased’ preferences and getting away with it. Abusive members are less likely to get away with their abuses to the extent that the reasoning behind their preferences is well-scrutinized.

If ‘harmony’, of the sort required by Conditions Four and Five, obtains among members’ preferences, then the amount of scrutiny of members’ preferences in the first-order network is increased. The second-order network increases members’ opportunities to scrutinize other members’ preferences, and ‘harmony’ ensures that the additional
scrutiny that it enables amounts to scrutiny of parties’ preferences in the first-order network, and not merely in the second-order network. Whenever parties to the second-order network scrutinize one another’s ‘waiving preferences’, they will effectively be scrutinizing one another’s preferences qua parties to the first-order network as well, for, if ‘harmony’ is guaranteed, a flaw in the reasoning behind a party’s waiving preferences entails a flaw in the reasoning behind her preferences among concrete actions.

I have said that the second-order network increases members’ opportunities to scrutinize other members’ preferences. By having ‘opportunities to scrutinize others’ preferences’ here, I really mean enjoying discursive entitlements that others should lay bare the reasoning behind those preferences, and heed one’s criticisms of that reasoning. The first-order network only entitles a given member S to scrutinize the voiceable preferences of those of whom S is demanding certain actions, and those who are demanding certain actions of S. It does not entitle S to scrutinize the preferences of those who are making demands of others, or whose preferences oppose the demands that S makes of others. The second-order network does permit this. The second-order network increases the amount of scrutiny of preferences that is going on, and if ‘harmony’ is established, this will all amount to scrutiny of first-order entrusted parties’ preferences, which are the preferences that influence members’ concrete choices. Equality of weighted desire-types for all in G, qua parties to the first- or the second-order network is useful as a step toward achieving ‘harmony’, and hence to increasing the scrutiny of first-order entrusted parties’ preferences.

(It will not be entirely clear yet why, as I have asserted above, a flaw in the reasoning behind a party’s waiving preferences entails a flaw in the reasoning behind her preferences among concrete actions. My argument in subsection 3.3.3 will explain this entailment.)

Greater scrutiny of first-order entrusted parties’ preferences would be to the advantage of all. But why would it be to the advantage of an individual member S of G to go to the bother of scrutinizing fellow members’ preferences? S wouldn’t have any trust-based obligation to scrutinize anyone’s preferences, only to defend the reasoning behind her own preferences when challenged. I have already alluded to one motive for scrutiny: by correcting defects in the reasoning behind another member’s preferences, S can potentially bring them into better alignment with her own. If first-order entrusted party U prefers to φ and S prefers that he not-φ, it might be worth S’s while to look for faults in U’s reasoning. But S’s motive for scrutiny here does not interest us, because
the discursive entitlements that facilitate S’s scrutiny do not depend on the second-order network; S can demand that U explain his reasoning because S has a voice in U’s decision qua first-order entrusting party.

Of some interest is the possibility that, if S could not change U’s preference, she would scrutinize the preferences of those first-order entrusting parties who preferred that U should φ, in the hope of changing all of their preferences, and so permitting U to yield to S’s preference without violating anyone’s trust. This strategy will be more promising if few of U’s partners preferred that he should φ, and if the terms of their trust relationships were such that U would significantly strain S’s trust if he did not have conflict excuses for resisting her preferences.

But even if, as would normally be the case, S had no hope of changing any particular decision of a first-order entrusted party, S might hope to bring the preferences of other members of G into better alignment with hers in a general way, by scrutinizing those preferences that she disagreed with. If S managed to change a member T’s preference that U should φ to a preference that U should not φ, S would raise the probability that, if, at some point in the future, T qua first-order entrusted party faced a decision that was similar to U’s, T’s discretion-licencing preference would be in alignment with S’s preference. (And perhaps in that future decision S wouldn’t have the time or opportunity to scrutinize T’s preference.) This would give S a motive, albeit a fairly weak one, to engage in many disagreements (in the activity sense of ‘disagreement’) even when winning would have no foreseeable practical issue.

The final and perhaps most important motive for scrutinizing preferences is glory. Just as a person may earn a reputation for cleverness by finding errors in other people’s theoretical reasoning, S might do the same by finding errors in the reasoning behind her fellows’ voiced preferences.104 (S might also hope to discredit her rivals by this activity.) This incentive to scrutinize would exist even if the discovery of error would produce no expected change in anyone’s behaviour.

Now that I have explained what I think the second-order network would look like if it existed, it should be relatively easy to see why I think it would exist. The members of G would be members of a second-order network because this would increase the

104 S will earn gratitude as well as esteem if her scrutiny reveals foul play, or a misjudgement that would have resulted in a serious frustration of common desires, though I think that this is relatively unlikely to occur if her scrutiny is of the sort that is enabled by the second-order network.
extent to which first-order parties’ preferences were scrutinized, by allowing members to scrutinize other members’ preferences qua parties to the second-order network. It would allow them to scrutinize the preferences voiced by entrusting parties in the first-order trust network, even if those preferences did not apply to any decision that they were to make. This scrutiny, and the threat of it, would reveal and deter abuses, and give parties a greater incentive to reason carefully when forming their accommodation-obliging and discretion-licencing preferences. The presence of this second-order network would permit any two members of G to have third-party disagreements, if they had opposing voiceable preferences concerning how another member of G should act.

Entrusting parties in a mere uniform weighted desires network lack this reason to form a second-order network. They don’t have much incentive to scrutinize one another’s preferences because they are not themselves common entrusted parties, and such scrutiny won’t change their behaviour. And though it might reveal bias in their voiced preferences, such bias is not a serious concern anyway, for the common entrusted party’s behaviour will rarely be affected by the particular preferences voiced by entrusting parties anyway.

3.2.2 Explaining Third-Party Disagreement in Photograph

In this subsection I will consider first whether my hypotheses concerning the base conditions for fully connected networks, in conjunction with the Four Relationships Hypothesis, predicts that Bertie’s extended family members can disagree with each other about whether he should send the photographs of Pete Seymour to the press. I will also touch on the question of why the conditions for third-party disagreement between Penny’s mother and aunt in Marriage (Imperatives) might be met.

My assumption that our intuitions about disagreement are accurate means that, when we judge that people in a described case disagree, we recognize that their circumstances are of a sort in which the social settings for disagreement would normally obtain. So if my account of the base conditions for dual-order fully connected networks is correct, and if I can make a plausible case that these conditions would obtain in Photograph if it were a real-world case, I will have explained our intuition of disagreement, given the assumption.

I contend that the base conditions for a dual-order fully connected network would indeed probably obtain in that case. First, the Seymours desire to see their family’s traditions honoured, and its prosperity and social status promoted. Different family
members will care more about some of these ends than about others. Thus they have common desires which they cannot serve, in a manner satisfactory to all, simply by following a common set of rules. Second, it is very plausible that Bertie is not the only Seymour who can satisfy or frustrate these common desires through his actions. Anyone could tarnish the family’s honour by involving herself in a scandal, or revealing the family's embarrassing secrets; anyone could violate the family's traditions; many in the family could damage their heirs’ chances of marrying well by squandering money, or engage in competition damaging to the family businesses. So it is very plausible that all or most adult Seymours would be entrusted parties in uniform weighted desires networks involving the other adult Seymours.

Thus there would plausibly be at least a subset G of Seymours such that the base conditions with respect to G would be conducive to the formation of a first-order fully connected network among G, and, if my argument above is correct, a second-order network entitling members to voices in one another’s waiving choices qua members of the former network. This second-order network would be useful, because it would facilitate more extensive scrutiny of members’ preferences. And it might very well permit Cousin Agnes and Aunt Augusta to disagree in preference about whether Bertie should send the photographs to the press.

An explanation of our intuitions about *Marriage (Imperatives)* will begin along the same lines as my explanation of our intuitions about *Photographs*, but it could be completed in more than one way, and I’m not sure which of these is best. We could say, to start with, that there is plausibly a fully connected trust network comprising Penny, her mother, and some subset of Penny’s extended family. Penny’s family is clannish like the Seymours; its members have common ends which all members could potentially frustrate, so forming a fully connected network would be useful to them as a means to discourage such frustrations. Greater scrutiny of voiced preferences would serve their interests too, so forming a second-order network would also be to their advantage.

Complications arise when we try to account for the presumed disproportionate power that Penny’s mother has over her. In general we should expect a trust network to be less appealing to its members, and so less likely to exist, if it does not give them a degree of conventional influence in one another’s choices that reflects their underlying power relationships. A fully connected network tends to constrain everyone’s choices in roughly the same way, not in a way that especially satisfies the desires of the powerful—or rather, not in a way that tends to ensure that, for all network members S,
T, and U such that S has power over U and T does not have power over U. U’s choices are constrained in a way that tends to satisfy S’s desires more than T’s. (If there are a few members who have power over almost everyone, this may influence the initial terms of the trust network: the desire-types that are counted as weighted might be skewed in order to constrain behaviour in ways that serve the interests of the powerful. But this is not presumably relevant to Penny’s case: we don’t assume that Penny’s mother is the matriarch of her extended family.) Up to a certain point, I think that this is an unavoidable feature of fully connected networks—it’s unavoidability explains, according to fiduciary non-cognitivism, why (as is plausibly the case) morality is a more irksome constraint on the powerful than on the powerless. But in Penny’s case it would be a bit surprising if everyone in Penny’s extended family had a voice in Penny’s choice of spouse that was equal in strength to that of Penny’s mother.

I offer some suggestions in Appendix D as to how Penny’s mother’s disproportionate power might be reflected in the arrangements of trust relationships among Penny’s family. In the meantime, I have found it convenient to use Penny’s case as a source of examples throughout the next section. To avoid the complications raised by Penny’s special relationship with her mother, I shall rely on a variation of the case which only involves Penny’s extended family. We may presume that an ordinary dual-order fully connected network exists among Penny’s family in this case, granting Penny’s relatives voices in her marital choice.

3.3 ‘Harmony’ and the Reaction Problem

I will now finally address the problem of how Condition Five of the Four Relationships Hypothesis can be satisfied. My solution is unfortunately rather radical, but it will allow me to solve another problem as well, which I call the Reaction Problem. The latter problem seems to provide better motivation for my radical solution than the former, so I will begin by discussing the Reaction Problem, then show how its solution provides the means for the terms of entrusted concern relationships to guarantee ‘harmony’ between a person’s voiceable preferences among concrete actions and her preferences among waiving choices, as required by Condition Five.

Suppose that (i) Penny has a pre-trust preference to marry Mr. Beau and a discretion-licencing preference not to marry him, (ii) Penny’s uncle, Arthur, has an accommodation-obliging preference that Penny should not marry Beau; (iii) if Penny goes ahead and marries Beau anyway, Uncle Arthur will avenge this violation of his
trust by disinheriting her; and (iv) Penny’s resulting poverty would embarrass the family. Penny’s cousin, Andrew, thinks that a marriage between Penny and Beau (a talented and promising young man from a family whose reputation has recently been marred by scandals) would improve the family’s standing in the long run, assuming Arthur did not exact his reprisal. Thus, on this assumption, Andrew would have an accommodation-obliging preference that Penny should marry Beau. But, since Arthur will condemn Penny to scandalous poverty if she marries Beau, Andrew’s preference, based on his weighted desires, is that Penny should not marry Beau.

Now it seems to me that the preference we would expect Arthur to voice, in a disagreement with Penny or Andrew or any other member of Penny’s extended family about whom Penny should marry, is not this second preference, but the former—the preference that he would have if he did not take into consideration Arthur’s reaction to Penny’s breach of trust. Suppose that Arthur said, ‘I’d prefer that Penny didn’t marry Beau: his family is in disgrace’, and Andrew replied, ‘Yes, I’d also prefer that Penny should reject Beau’s suit, because if she married him you would condemn her to scandalous poverty, though I think that the marriage would otherwise have improved our family’s standing.’ This reply sounds odd or perhaps ironic, not like the sincere voicing of a preference. Intuitively, Andrew and Arthur really disagree about whether Penny should marry Beau. It seems to me that Arthur would form those preferences he might sincerely voice without taking account of the likely reactions of the other entrusting parties, as such, to Penny’s marrying or not marrying Beau. Vaguely put, it seems that he would consider the marriage ‘on its own merits’.

The Reaction Problem, as it relates to entrusting parties’ preferences, arises when we take these appearances at face value, as I shall. It is the problem of explaining how it could be the case that entrusting parties in uniform weighted desires networks, and other similar trust networks, rationally hold voiceable preferences that they could only hold because they are disregarding facts about other entrusting parties’ reactions (qua entrusting parties) to the entrusted party’s actions. This defies the characterization of accommodation-obliging preferences that I have given previously, according to which they are simply preferences based on a certain subset of the entrusting party’s desires, those specified as weighted by the terms of her entrusted concern relationship.
3.3.1 The Admissible Grounds Hypothesis as it Applies to Entrusting Parties’ Preferences

My solution to the problem is simple. I shall present a revised characterization of accommodation-obliging preferences, according to which a preference, based solely on weighted desires, can fail to be accommodation-obliging if it does not omit to take account of certain grounds. I call this the Admissible Grounds Hypothesis as it relates to accommodation-obliging preferences (I will later extend it to apply to discretion-licencing preferences). I will use a stipulative notion of ‘grounds’ for preferences in my characterization of this hypothesis, which is supposed to broadly approximate our commonsense notion. The stipulative notion is defined below. The first definition, that of a ground relative to a probability function, is fundamental; the latter two depend on it.

**Ground relative to a probability function.** Proposition \( q \) is a ground for an agent \( S \) to hold a preference that \( p \) based on a set of desire-types \( W \) relative to a probability function \( Pr \) iff, for \( S \), the rational expected utility in terms of \( W \)-type desire-satisfaction of \( p \) on the supposition that \( q \) would be greater than that of \( p \) on the supposition that \( \neg q \), if \( S \)’s probability function were \( Pr \).

**Ground ‘simpliciter’.** Proposition \( q \) is a ground simpliciter (i.e. not relative to any specified probability function) for an agent \( S \) to hold a voiceable preference that \( p \) based on a set of desire-types \( W \) iff (i) \( q \) has some positive objective probability, and (ii) \( q \) is a ground for \( S \) to hold a preference that \( p \) based on \( W \) relative to the objective probability function.

**Ground of an agent.** Proposition \( q \) is one of agent \( S \)’s grounds for holding a preference that \( p \) based on a set of desire-types \( W \) iff (i) \( S \) has some positive credence that \( q \), and (ii) \( q \) is a ground for \( S \) to hold a preference that \( p \), based on \( W \), relative to \( S \)’s probability function (i.e. her credal state).

According to the Admissible Grounds Hypothesis, an entrusting party’s accommodation-obliging preference that \( p \) is a preference that is based on all and only her weighted desires, taking into account all and only a certain subset of her grounds for preferring that \( p \) and \( \neg p \)—her admissible grounds, as I shall call them. (If a ground is not admissible, then it is inadmissible.) An entrusting party’s admissible grounds are those of her grounds that fall under a set of ground-types, which I will call her admissible ground-types. These are specified by the terms of her entrusted concern relationship.

This proposal, once refined, will allow us to solve the Reaction Problem. How is it that entrusting parties in uniform weighted desires networks can rationally hold voiceable preferences that they could only hold because they are disregarding facts.
about other entrusting parties’ reactions to the entrusted party’s actions? Because propositions concerning reactions of entrusting parties as such to the entrusted party’s actions are not admissible grounds for accommodation-obliging preferences.

I will present some refinements of the Admissible Grounds Hypothesis shortly. Before doing so, I will state an assumption that I will be making concerning admissible ground-types. This is a ‘virtuous’ assumption, because by making it, I avoid committing to claims about admissible grounds beyond those I need to make in order to explain the appearances that I adduced above in support of the Admissible Grounds Hypothesis. It roughly concerns the extent to which mutually inconsistent grounds with the same ‘subject matter’ can differ in whether they do or do not fall under admissible ground-types. Having thus clarified my commitments, I will defend a hypothesis concerning what the process of omitting to ‘take into account’ grounds actually involves.

The assumption about admissible ground-types that I shall make has two parts. First, if any proposition falls under an admissible ground-type, then so does its negation; and if a complex proposition, composed of atomic propositions joined by logical connectives, falls under an admissible ground-type, then so do all the variations on that complex proposition that we get by negating different combinations of its components. Second, if any proposition whose content contains gradable predicates falls under an admissible ground-type, then so does any otherwise identical proposition whose content contains gradable predicates of the same sort but which pick out different values, or ranges of values. Thus if Mr. Beau has high social status falls under an admissible ground-type, so does Mr. Beau has low social status.

This is a complicated assumption, and one that doesn’t serve my broader ends, for it makes explaining ‘harmony’ more difficult, and is unnecessary for my metaethical theory. Beyond the fact that working within its constraints is a way to avoid ad hoc commitments, I make this assumption because it rules out certain possibilities that I find counterintuitive. To wit, it would allow that certain specifications of admissible ground-types might ‘skew’ entrusting parties’ accommodation-obliging preferences in certain directions, not just by forcing them to disregard certain subject matters, but by forcing

105 What has just been said of atomic propositions applies also to the ‘predications’ of quantified variables, regardless of the scope of the quantifiers. Suppose that the following proposition falls under an admissible ground-type: For all x such that x is a suitor of Penny, there is some y such that (y has a knighthood and y is related to x) if and only if x is not Mr. Beau. This being so, the variations on the proposition that we get from negating its various predications (e.g. y has a knighthood, y is related to x) also fall under an admissible ground-types.
them to disregard all and only those propositions on certain subject matters that favoured certain preferences.

For instance, take the following proposition:

(1) If Penny marries Mr. Beau, the family will lose money.

Suppose that Cousin Andrew has a weighted desire that his family should be richer, and believes that Penny’s marrying Beau will increase his family’s wealth. And suppose that, contrary to my second postulation, (1) is the only ground concerning the effects of Penny’s marrying Beau that makes a prediction about the consequences of the marriage for the family’s wealth that falls under an admissible ground-type. The proposition that Penny’s marriage will make the family richer is thus not among Andrew’s admissible grounds for preferring that Penny marry Beau. However, (1) is among Andrew’s admissible grounds for preferring that Penny not marry Beau, for Andrew will have some credence in (1). So his desire that the family should be richer will end up supporting a preference that Penny not marry Beau, even though he believes that the marriage will make the family richer. If he ended up with an accommodation-obliging preference that Penny should not marry Beau, and were asked to give the reasons for his preference, we might expect him to cite his fears that the marriage would make the family poorer, even though he believed the contrary proposition. This strikes me as odd. To the extent that accommodation-obliging preferences are skewed in this way, the correctness of calling them preferences seems doubtful (even on my admittedly elastic usage); and I am not sure that non-cognitive ‘clashes’ involving thusly skewed preferences would elicit intuitions of robust non-cognitive disagreement.

Given my present assumption, however, this sort of skewing should not happen, or at least not in cases like the one I have just described. If (1) fell under an admissible ground-type then so would the following (among others):

(2) It is not the case that, if Penny marries Mr. Beau, the family will lose money.
(3) If Penny does not marry Mr. Beau, the family will lose money.
(4) If Penny marries Mr. Beau, it is not the case that the family will lose money.
(5) If Penny marries Mr. Beau, the family will gain money.

The first part of my assumption, concerning negations of propositions falling under admissible ground-types, ensures that (2)-(4) fall under admissible ground-types. The second part, concerning gradable predicates, ensures that (5) falls under an admissible ground-type. Since (5) is an admissible ground for Andrew, the strength of his
preference will reflect his desire that his family should be wealthier and his belief that
the marriage will make it wealthier, just as we would expect.

I have said that I have made this assumption because I find the sort of ‘skewing’
that it apparently prevents counterintuitive. I should admit, however, that ‘skewing’ of a
not-too-dissimilar kind is possible even given my assumption. The terms of entrusted
concern relationships can, by craftily making certain ‘subject matters’ inadmissible,
block off certain paths of inference that would have supported beliefs in admissible
grounds, while permitting inferences supporting inferences in their negations. (My
proposal in 3.3.3 about how ‘harmony’ would be guaranteed might amount to an
example of this sort of counterintuitive skewing.) It is not clear to me whether these
latter kinds of ‘skewings’ are as counterintuitive as the kind I described above.

Now I shall make a proposal concerning the adjustments an entrusting party must
make to the representation of the world that she entertains when she disregards a
proposition because it is inadmissible. According to my proposed interpretation of the
original example of Penny and Mr. Beau, Cousin Andrew must disregard the following
proposition, which I will label $q$, when forming his preference: *Uncle Arthur will
disinherit Penny if and only if Penny marries Beau.* She must do so because
propositions concerning reactions of members of the familial trust network as such to
the entrusted party’s actions are not admissible grounds. But it is not clear why this
means that $q$ must be disregarded. $q$ does not imply that Arthur’s action would be the
reaction of a network member as such to an entrusted party’s actions. Its truth could
have any number of explanations. Perhaps the following is true: Arthur only plans to
leave part of his estate to Penny because he lacks children of his own; if Penny marries
Beau, this will put Arthur himself in a mood to marry, with the result that he will
produce offspring and make them his sole heirs. If this were the reason why Andrew
believed $q$, we would not expect him to disregard $q$ when forming his preference. Thus I
don’t want to hypothesize that $q$ falls under an inadmissible ground-type. I want to
hypothesize that Andrew would disregard it because of what he knows about the
reasons why Arthur would disinherit Penny.

I propose that Andrew would have to disregard $q$ because $q$ is *derivatively*
inadmissible—because he could not explain why he believed it without citing beliefs or
credences in propositions that fell under inadmissible ground-types, or *fundamentally*
inadmissible propositions. To ‘explain’ here is to give a *justifying* explanation, or—to
use a less obviously normative term—an explanation of the sort that would establish, or contribute to establishing, that his belief in \( q \) was rationally coherent.

Now suppose that Andrew’s belief in \( q \) is supported by exactly two distinct chains of inference. He has some credence in the proposition that Arthur would disinherit Penny if she married Beau as a reprisal for her abuse of his trust, but also has some credence in the proposition that he would disinherit her because her marriage would put him in the mood for marriage himself, with the result that he would have children and make them his sole heirs. I propose that in this case \( q \) is inadmissible insofar as Andrew’s belief in it is based on the ‘reprisal’ inference, because this cites fundamentally inadmissible propositions concerning the reactions of network members as such; and that \( q \) is admissible insofar as it is based on the ‘mood’ inference. When forming his preference, Andrew must not take the support his belief in \( q \) gets via the ‘reprisal’ inference into account, and must ‘hypothetically’ proportion his credence in \( q \) to the support it gets via the ‘mood’ inference alone. (The ‘part’ of his credence in \( q \) that is supported by the latter inference alone is, I shall say, his admissible credence in \( q \).) In general, I propose that, roughly, an entrusting party’s preference is accommodation-obliging if and only if it is based on all and only her weighted desires, and on her grounds just to the extent that they are admissible.\(^{106}\)

3.3.2 The Hypothesis Applied to Discretion-Licencing Preferences

I think that there is a version of the Reaction Problem that pertains to entrusted parties’ preferences (whether or not it also pertains to entrusting parties’). It has nothing in particular to do with networks of trust relationships. I might have mentioned it in Chapter 1 and used it to support a postulation of admissible and inadmissible grounds in that chapter, but it seems an obscure and untroubling issue when considered in isolation from other instances of the Reaction Problem.

The extent to which an entrusted party strains his partner’s trust has practical significance, and it can affect the extent to which his weighted desires are satisfied. If John yields in Football (*Mother and Son*), he relieves strain on his mother’s trust,\(^{107}\) and increases his scope to resist her preferences without violating her trust in the future.

---

\(^{106}\) This biconditional is not to be interpreted as a strict definition; it is not meant to answer every question we might have about conditions for accommodation-obligingness, such as whether very irrationally formed preferences can be accommodation-obliging.

\(^{107}\) Or discharges ‘debt’, as discussed in Appendix 1. I ignore the distinction between strain and debt in what follows.
Suppose that John wants to stay with a friend who lives on the coast in the summer and learn how to surf, because he thinks that he will enjoy surfing. He knows that his mother would rather he didn’t, because surfing is dangerous—indeed more dangerous than football. John knows that if he strains his mother’s trust by playing football in the winter, he won’t be able to resist her preference that he not surf in the summer without violating her trust. She will forbid him from staying with his friend in the summer if he doesn’t relieve the strain he has already placed on her trust, by being acquiescent in the meantime—she will feel free to do this, in accordance with their mutual understanding, because he will have reached the limit of his discretion. Thus, all things considered, John doesn’t want to play football in the winter, because he wants to relieve his mother’s trust so that he may learn to surf in the summer. The trouble is that learning to surf would satisfy his weighted desires—specifically his desire for enjoyment. Since he wants to forbear from playing football as a means to bring it about that he surfs, and since this is explained almost entirely by his weighted desire for enjoyment, John has a discretion-licencing preference not to play football. But if he has a discretion-licencing preference not to play football, then he does not relieve strain on his mother’s trust by forbearing from playing football. He would simply violate her trust by playing football, because he lacks discretion to do so.

Now, I think, we encounter a paradox: since John’s forbear from playing football would no longer serve its purpose, John would no longer want to forbear from playing football, and so he would once again have a discretion-licencing preference to play football. So once again, if he refrained from playing, he could bring it about that he surfed in the summer. But he would want to do this for reasons of enjoyment, and so he would have a discretion-licencing preference not to play football. Then his forbear from playing football would not serve its purpose, and he would instead prefer to play football. And so on.\textsuperscript{108}

If this is a genuine paradox, then we should abandon any theory that entails its possibility. Even if it is not a paradox, it is bizarre to suppose that John would be unable to relieve strain on his mother’s trust just because doing so was a means to satisfying his weighted desires rather than his unweighted desires. The problem arises because, on my

\textsuperscript{108} It is not clear to me whether the option of yielding to his mother’s preference on the football question by verbally giving her his assurance that he will not play football, rather than by merely not playing football, provides a way for John to relieve strain on his mother’s trust. But if it does, we could preserve the apparent paradox by stipulating that yielding in this fashion is not an option for John in his circumstances.
present characterization of discretion-licencing preferences, John must take account of his beliefs about his mother’s reactions, qua entrusting party, to his actions when forming discretion-licencing preferences. It is a version of the Reaction Problem.

We can avoid these unhappy predictions by extending the Admissible Grounds Hypothesis so that it applies to discretion-licencing preferences. We make the following two postulations. First, an entrusted party B’s discretion-licencing preference that \( p \) is a preference that is based on all and only his weighted desires, taking into account B’s grounds for and against this preference only to the extent that they are admissible. A ground of B’s is not admissible (it is ‘fundamentally inadmissible’) if it does not fall under an admissible ground-type, the set of which is specified by the terms of his entrusted concern relationship. A ground is also not admissible to the extent that B’s belief in it depends on inferences from beliefs in propositions that are fundamentally inadmissible (to that extent, it is ‘derivatively inadmissible). Second, propositions concerning the reactions of an entrusting party as such to her partner’s actions are (typically, and in the relationship between John and his mother at any rate) fundamentally inadmissible grounds for the latter’s discretion-licencing preferences.

3.3.3 Predicting ‘Harmony’

I will shortly offer an explanation as to how ‘harmony’ might be realized. But for this explanation to work I need to make a somewhat ad hoc postulation about how waiving works. When I introduced the idea of waiving in 2.2.2, I said that it was a conventional action, but I did not say what kind of action it was. I did not say, for instance, that it was a speech act that an entrusting party performed in the presence of her partner. Now I need to postulate that it is not an action of this sort. Rather it is the act of adopting a certain kind of conventional ‘stance’ which affects an entrusted party’s obligations, and places certain obligations on the waiving party. The defining properties of the waiving stance are as follows. For all \( \phi \), an entrusting party A can waive her demand that her partner B should \( \phi \) just in case A voiceably prefers that B should \( \phi \). If A adopts a stance of waiving with respect to B’s \( \phi \)-ing, then, as long as she continues in this stance:

(i) if B not-\( \phi \)s despite lacking a voiceable preference to not-\( \phi \), then B does not violate A’s trust by doing so (even if B does not know that A has waived); and

(ii) if a suitable occasion arises, and if A thinks that it will make a difference to whether or not B \( \phi \)s, A must communicate to B that she has waived with respect to B’s \( \phi \)-ing.
Note that I have dropped my earlier postulation that an entrusting party can only waive if she and her partner voiceably prefer the same option, or her partner lacks a voiceable preference. An entrusting party can assume a waiving stance regardless of her partner’s preference, but her waiving will only have practical significance if her partner does not disprefer her preferred option. The obligation stated in (ii) is of course purely conventional, like the other norms governing entrusted concern relationships; the penalty for violating it is just a possible diminution in one’s esteem or one’s reputation for being a cooperative entrusting party. I will assume that the waiving stance is one that an entrusting party may adopt or forswear at any time without any necessary impropriety. However if she has waived out of deference to the accommodation-obliging preferences of a second-order entrusting party, then her abandoning the stance will presumably strain that party’s trust.

Although this postulation concerning waiving is somewhat ad hoc and burdensome, it is not one I am making purely so that I can explain how ‘harmony’ might be realized. In some dual-order fully connected networks, like the moral trust network, and perhaps the trust network to which Penny and her extended family belong, it is implausible to suppose that waiving should necessarily involve the performance of a speech act in the presence of an entrusted party. Suppose that two moralists disagree and argue about whether another moralist on the far side of the world should φ. It is not plausible that one moralist expects that if he wins the argument, the other moralist will traverse the globe to meet the third-party moralist and perform a speech act in his presence, or even perform the same act by email or telephone. Likewise, if Uncle Arthur wins an argument with Cousin Andrew about whom Penny should marry, Arthur would not necessarily expect Andrew to go and perform a speech act in Penny’s audience, if he doesn’t think that it will affect her choice either way.

We can explain the ‘harmony’ between preferences among waiving choices and preferences among concrete actions required by Condition Five of the Four Relationships Hypothesis, by positing that entrusted concern relationships in the second-order network would make (virtually) any grounds that would interfere with this ‘harmony’ inadmissible. And since we have independent support for the view that there can be admissible and inadmissible grounds for entrusting and entrusted parties’ preferences in some entrusted concern relationships, we can do so at a smaller cost to the plausibility of the Four Relationships Hypothesis. Importantly too, we do not have to posit ‘harmony’-preserving admissible ground-types merely in order to get my theory
of third-party non-cognitive disagreement to explain the facts. The ‘harmony’ we predict by doing so would serve the interests of parties to fully connected networks by facilitating more extensive scrutiny of preferences. (These considerations notwithstanding, my explanation as to how ‘Harmony’ would be guaranteed will require me to make some quite costly commitments.) I’ll now expand upon these points.

First, let’s consider more specifically which ground-types for parties in second-order entrusted concern relationships need to be inadmissible if harmony is to be achieved. Suppose that S and T have a third-party disagreement $\Delta_{(S,T,U,\phi)}$ about whether or not U should $\phi$. U prefers to $\phi$; S prefers, qua U’s partner, that U should not-$\phi$, and T prefers, qua U’s partner, that U should $\phi$. Let’s suppose that S, T, and U share the four relationships in virtue of which S and T disagree about whether U should $\phi$ because they are parties to a dual-order fully connected network N. Thus all parties in all relationships have the same weighted desire-types W.

It must be possible for S to influence U’s actions via the following route: (i) S voices her preference to T that U should not-$\phi$; T is moved to waive by his desire not to strain S’s trust; this increases (minutely) the chance that all entrusting parties who prefer that U should $\phi$ waive; if they do all waive, U will then have to not-$\phi$ in order to avoid straining S’s trust, and will be motivated to do so by his accommodative desires. Of course, unless N is quite small, S will never voice her preference to T that U should not-$\phi$ out of a forlorn hope of changing U’s behaviour in the way just described. She will normally do so as part of an effort to change T’s own preferences by challenging him to explain why the reasoning behind her preferences does not give him preferences for the same option, with the end of changing T’s own behaviour, or for the glory of defeating T in argument. But it is because of the availability of this unlikely means of influencing U’s actions that S will prefer that T should waive when she prefers that U should not-$\phi$, and T and U prefer that U should $\phi$. (So strain must matter, albeit nominally, in all networks of trust permitting third-party disagreements; entrusted parties must have some motivation to avoid it.)

Some ground-types are (fundamentally) inadmissible in the first-order relationships $E(S,U,\Delta_{(S,T,U,\phi)})$ and $E(T,U,\Delta_{(S,T,U,\phi)})$. These include what I will call ‘N-grounds’, which are a class of propositions concerning N. N-grounds include propositions concerning the conventional roles and actions that people occupy and
perform qua members of N, the consequences of occupying and performing these roles and actions, the conditions of trust among members of N as such (e.g. how much strain they have accumulated); the reactions of members of N to the straining, honouring, or violation of their trust by other members of N as such; the trust-based desires of members of N as such, the voiceable preferences of members of N; and so on. Colloquially, members of N are to ‘forget all about N’ when forming their preferences concerning the actions of entrusted parties in N’s first-order entrusted concern relationships.

I don’t think that the terms of the relationships in dual-order fully-connected networks in general need be as sweeping as those of N are in ruling grounds relating to the trust network itself inadmissible. (Indeed, the moral trust network is a dual-order fully connected network, and it allows members to form judgements about the moral propriety of being morally hypocritical and ‘judgemental’. If N were the moral trust network, propositions about being hypocritical and judgemental would be inadmissible.) But I am merely providing an idea of how ‘harmony’ in dual-order fully connected networks is possible, and stipulating that all N-grounds are inadmissible in N’s first-order relationships makes this simpler.

There is one class of propositions that are not N-grounds but which I shall also have to postulate would be fundamentally inadmissible—for all members of N in fact, regardless of their roles. I will give the details of this postulation at a later point, when I am ready to explain why it is needed.

To predict ‘harmony’ among S’s preferences, we can hypothesize that the admissible ground-types for S in \( E(S,T,\Delta_{(S,T,U,\phi)} \) are equal to the admissible ground-types for S in \( E(S,U,\Delta_{(S,T,U,\phi)} \), plus some additional types of N-grounds. These N-grounds include a proposition, which I will call the N-conclusion, and some premises in an argument implying that proposition, which I will call N-premises. In the toy example of such an argument that I set out below, the N-conclusion is the proposition that \( U \text{ will not-\( \phi \) if and only if } T \text{ waives} \). The N-conclusion is supposed to be, under virtually all circumstances, the only N-ground that S needs to have in mind when she reasons about her accommodation-obliging preferences in \( E(S,T,\Delta_{(S,T,U,\phi)} \); the other admissible N-

\[109\] Propositions concerning acts of yielding by entrusted parties N, where the yielding is done through the performance of concrete actions rather than through giving assurances of acquiescence, are admissible. But they are only admissible if they do not imply that the concrete actions constituted yielding by N-members.
grounds merely provide an inferential route via which she can form a positive admissible credence in the N-conclusion. The content of the N-conclusion is supposed to be such that, if it is the only N-ground S has in mind (i.e. that influences her reasoning), and if it is admissible, and if S has a positive credence in it, ‘harmony’ will obtain: S will prefer in $E(S,T,\Delta_{(S,T,U,\varphi)})$ that T should waive his demand that U should $\varphi$ when and only when S prefers in $E(S,U,\Delta_{(S,T,U,\varphi)})$ that U should not $\varphi$. It should be evident that the content of the N-conclusion from my example has this property at least.

The argument whose conclusion is the N-conclusion must be a means by which B can always find some epistemic support for the N-conclusion—that is, the essential premises in the argument must be such that S always has some positive credence in each, if she is rational. The premises that are not deducible from other premises must be ones that S can believe because they are implied by the ‘standardized’ terms of the relationships in N (i.e. the terms that all relationships in N have in common), or because they are implied by assumptions S must make if she is to be a trusting and committed member of N (e.g. the assumptions that U and other members are minimally trustworthy, have appropriate trust-based motivations, and know the terms of the trust relationships to which they are parties qua members of N), or because believing them does not require any obvious inference. This last category includes propositions that are platitudinous; or else are ‘directly’ supported in some way by S’s present observations, by her intuitions, or by introspection. These platitudinous or ‘directly-supported’ propositions must fall under admissible ground-types if they are to play a role in the argument. Not all of the premises in the argument must be N-grounds, but all must fall under admissible ground-types, and so must the N-conclusion itself.

There are other admissible N-grounds apart from the N-conclusion and the N-premises needed to support it. For, as the reader will recall, if a proposition $q$ falls under an admissible ground-type, then so must all the variations on $q$ that we get when we negate $q$, or modify $q$ by negating combinations of its atomic-proposition components or altering the values of its gradable predicates. I’ll call these variations proposition-variants. The admissible N-grounds consist exclusively of the N-conclusion, the N-premises supporting it, and the proposition-variants of these. One of the challenges of explaining ‘harmony’, on my assumptions, is to show that there is no argument from the proposition-variants of the N-premises to a proposition-variant of the N-conclusion which is inconsistent with the N-conclusion, where this argument will sometimes lend
greater credence to the N-conclusion-variant than the N-premises do to the N-conclusion.\textsuperscript{110} This would result in a failure of ‘harmony’.

Below I offer an example argument to illustrate what the argument for the N-conclusion might look like. The example is merely illustrative. It would not serve the purposes of any real trust network. It ignores the realistic possibility that U might be unaware of what the other members of N prefer that he should do, and of whether those who prefer that he should φ are willing to waive. Taking account of these possibilities would make the argument quite a bit more complex. I have also not stated the argument very rigorously,\textsuperscript{111} because this would hinder readability, and because I do not aspire at this stage in my theorizing to a high degree of precision. A final cause for wariness about my argument is that it only provides a route by which S can form a positive admissible credence in the N-conclusion if we make the following very controversial assumptions about degrees of belief: first, that (positive) infinitesimal credences are possible; second, that it is irrational to have a credence of zero in any contingent proposition (a property of ‘regular’ probability functions);\textsuperscript{112} and third, that infinitesimal credences can affect what one may rationally prefer.

The argument consists of two distinct arguments for lemmas L1 and L2, which respectively imply the right-to-left and left-to-right parts of the biconditional C. A ‘pro-φ entrusting party’ (see P4-P11) is a member of N who has an accommodation-obliging preference that U should φ.

\begin{itemize}
  \item P1 U has some trust-based desire not to strain S’s trust
  \item P2 If U has some trust-based desire not to strain S’s trust, then U will not strain S’s trust
  \item P3 U will not strain S’s trust [P1 & P2]
  \item P4 If all pro-φ entrusting parties in N waive and U φs, then U will strain S’s trust
  \item P5 If all pro-φ entrusting parties in N waive, then U will not-φ [P3 & P4]
  \item P6 For all x, if x is a pro-φ entrusting party in N, and x is not T, x will have some positive credence that there is some y such that y is a member of N who prefers that x should waive
\end{itemize}

\textsuperscript{110} The sort of credence I have in mind here is what I call below a ‘difference-making credence’.

\textsuperscript{111} It mixes ordinary language and formalism, glosses over changes in tense, and relies on some unstated premises.

P7 For all \( x \), if \( x \) is a pro-\( \phi \) entrusting party in \( N \), and \( x \) is not \( T \), and \( x \) has some positive credence that there is some \( y \) such that \( y \) is a member of \( N \) who prefers that \( x \) should waive, then \( x \) will have some positive credence that there is some \( z \) such that \( z \) is a member of \( N \) and \( x \) would strain \( z \)’s trust by not waiving.

P8 For all \( x \), if \( x \) is a pro-\( \phi \) entrusting party in \( N \), and \( x \) is not \( T \), and \( x \) has some positive credence that there is some \( y \) such that \( y \) is a member of \( N \) and \( x \) would strain \( y \)’s trust by not waiving, then \( x \) will have some trust-based desire to waive.

P9 For all \( x \), if \( x \) is a pro-\( \phi \) entrusting party in \( N \), and \( x \) is not \( T \), and \( x \) has some trust-based desire to waive, then \( x \) will waive.

P10 All pro-\( \phi \) entrusting parties in \( N \) except \( T \) will waive [P6-P9]

P11 If \( T \) waives, then all pro-\( \phi \) entrusting parties in \( N \) will waive [from P10]

L1 If \( T \) waives, then U will not-\( \phi \) [P5 & P11]

P12 U will not violate \( T \)’s trust

P13 If \( T \) does not waive and U does not \( \phi \), then U will violate \( T \)’s trust

L2 If \( T \) does not waive, then U will \( \phi \) [P12 & P13]

C U will not-\( \phi \) if and only if \( T \) waives. [L1 & L2]

The conditionals in this argument are material conditionals. Thus a person might rationally believe one of them merely because she had a sufficiently low credence in its antecedent or a sufficiently high credence in its consequent, despite lacking any beliefs about correlations or relations of dependence between the truth of the antecedent and that of the consequent. However I am not interested in S’s beliefs in any of the conditional lines of this argument to the extent that they are based merely on unconditional credences in the antecedents or consequents. Credences in conditional lines arrived at in this way will not make any difference to S’s rational preference between \( T \)’s waiving and \( T \)’s not-waivering. To the extent that her beliefs in the conditional premises are explained in this way, she cannot coherently change her credences in their consequents or antecedents through modus ponens or modus tollens inferences involving those conditionals. To the extent that her beliefs in L1 or L2 are explained in this way, these beliefs will not affect her preference qua \( T \)’s partner. Except in extreme cases, S’s degree of confidence that \( T \) will waive or that U will \( \phi \) is beside the point when she is forming her preferences as to whether \( T \) should waive as a means to affecting whether U \( \phi \)s. I will only be interested in S’s credence in any of the conditional lines to the extent that it is based on what I will call her difference-making credences, such as her conditional credence in the consequent given the antecedent, or
her beliefs about the modal relationship between the truth of the consequent and that of the antecedent.\footnote{The N-conclusion is also to be be interpreted in the standard truth-functional way, which means that one might believe it merely because one believes the propositions on both sides of the biconditional, or disbelieves them both. What I say about being interested only in S's 'difference-making' credences in conditionals applies also to biconditionals such as the N-conclusion and its proposition-variants.}

Before I consider the sources of S’s justification for believing the premises of this argument, I want to say something about the overall plausibility of the argument from S’s point of view, and how we should expect it to affect her preference qua T’s partner. The argument for L1 is very weak because it relies on premises P2 and P9. P9 will presumably be very implausible, and increasingly so the larger N is. P2 will also be very doubtful in typical cases. Because of the implausibility of these premises, I shall need it to be the case that there is no argument with admissible premises and an admissible conclusion, that would undermine a difference-making credence in L1. As far as I can see, the only such argument would be one whose conclusion was the following L1-variant: \textit{If $T$ waive, then $U$ will $\varphi$}. Call this L1'.

Unfortunately, for all I have said so far, L1' does enjoy some support from a source that is admissible because it is not an N-ground. This is what I shall call the \textit{Contingency Platitude}, the proposition that if $p$ is a contingent proposition, there is a positive probability that $p$. Since L1' is itself a contingent proposition, and since it is implied by admissible contingent propositions (e.g. variants of the N-premises), the Contingency Platitude gives a (very small) degree of support to L1'. Of course it also gives a similar degree of support to L1. But this may not actually improve the plausibility of L1. The support L1 gets from P1-P11 may be infinitesimal, as will any support it gets from the Contingency Platitude; thus the combined support of both will remain infinitesimal. Since the Contingency Platitude will not increase the plausibility of L1, the best I can do is to avoid predicting that it would increase the plausibility of L1'. I shall thus postulate that some versions of the Contingency Platitude are fundamentally inadmissible for all members of N. I might say, for instance, that any version that implies that any N-ground has a positive probability of being true does not fall under an admissible ground-type.

If this postulation is true, then there are no admissible premises supporting L1'. This is because the line of support that P1-P11 gives to L1 concerns facts about whether pro-$\varphi$ entrusting parties have trust-based motivations to waive, and whether $U$ would
have trust-based desires not to strain S’s trust if they did waive. Any arguments for the L1-variant based on the P1-P11-variants will thus have to rely on the same sorts of facts. But if we confine our attention to these sorts of facts, we do not find any support for L1’. All pro-φ entrusting parties will have some trust-based motivation to waive. None has a trust-based motivation to not-waive. These facts give no support to the (otherwise very plausible) proposition that some pro-φ entrusting parties will not waive. If pro-φ entrusting parties do all waive, then U will have some trust-based motivation (qua party to E(S,U,Δ(S,T,U,φ)) at least) to not-φ, and no trust-based motivation to φ. So as long as S is considering these sorts of facts alone, she will not give difference-making credence to L1’. L1 is not wholly derivatively inadmissible; L1’ plausibly is.

This illustrates my general strategy for ensuring that the N-conclusion is the only distinctive admissible ground for S that will influence S’s preferences qua second-order entrusting party. We locate a set of questions or subject matters a consideration of which will robustly yield such a difference-making credence in the N-conclusion. Then we hypothesize that propositions concerning all other questions and subject matters that might lead S to have higher difference-making credences in propositions that contradict the N-conclusion would be fundamentally inadmissible, along with any that might lead her to believe that T’s waiving choice would have weighted-desire-satisfying properties beyond its effects on U’s choice. The more closely the ‘admissible’ questions and subject matters are confined so as to permit only the desired line of thought, the less likely is the occurrence of a deviation from harmony.

In contrast to the argument for L1, the argument for L2 is quite strong. At least, it is one that S must find somewhat plausible, as a committed party to N herself, and as someone who trusts U.114 There is no credible argument with admissible premises that will lower S’s difference-making credence in L2. This is, I think, just to say that there is no credible argument with admissible premises by means of which S could raise her difference-making credence in the L2-variant If T does not waive, then U will not-φ. I’ll

114 Even if S does not have a very high difference-making credence in L2, I find it plausible to suppose (i) that S must proceed as if L2 were a certainty, as an expression of her trust in U as a committed member of N; and (ii) that, if S’s voiced preference depended on her making this ‘trusting presumption’, this would not disqualify it from being accommodation-obliging. I might have to rely on this sort of idea more if I were offering a more rigorous explanation of ‘harmony’, on less forgiving assumptions. On the related idea that trust involves an attitude of optimism that ‘restricts … the kinds of inferences we will make about the likely actions of another’, see Jones, ‘Trust as an Affective Attitude’, 11.
call this conditional $L_2'$. Here are the deductive arguments for $L_2'$ make use of variants of the premises supporting $L_2$:

\[
\begin{align*}
P_{12}' & \quad \text{U will violate T’s trust} \\
P_{13}' & \quad \text{If T does not waive and U φs, then U will not violate T’s trust} \\
L_2' & \quad \text{If T does not waive, then U will not φ} \\
P_12 & \quad \text{U will not violate T’s trust} \\
P_{13}'' & \quad \text{If T does not waive and U φs, then U will not violate T’s trust} \\
L_2' & \quad \text{If T does not waive, then U will not φ}
\end{align*}
\]

An inference involving the first argument will fail to give S an admissible credence in $L_2'$ because $P_{12}'$ is derivatively inadmissible. $P_{12}$ is admissible only to the extent that it is entailed by S’s assumptions about U as a party to N. $P_{12}'$ is not supported by this route, and has no non-inferential support—its truth is not intuitive or directly observable or introspectable. An inference involving the second argument will similarly fail because $P_{13}''$ has no non-inferential support and is contradicted by the standardized terms of the relationships in N. (Similar arguments using variants of $P_{12}$ and $P_{13}$ to support $L_1'$ face similar problems.)

S should have a positive (and typically fairly high) admissible difference-making credence in $L_2$, and a positive (though perhaps infinitesimal) admissible difference-making credence in $L_1$, and no admissible difference-making credence in $L_1'$ and $L_2'$. Thus it will presumably be rational for her to prefer that T should waive when and only when she rationally disprefer an option of U’s that both T and U prefer. I say ‘presumably’ because, if she has very unusual desires, and those desires are weighted in N, failures of harmony are possible. For instance, if S has a basic desire that T should not waive, and this desire is weighted, then this might move her to prefer that T should not waive even when she prefers an option available to U that is incompatible with the option T and U prefer; and the limitations on admissible N-grounds that are present in my example would not prevent this.\(^{115}\)

I haven’t yet explained why S is justified in believing all of the premises of my argument for the N-conclusion, to the extent that she is. Premises P3, P5, P10, and P11

\[^{115}\text{It might be thought that if S had an unusual belief, such as that God would destroy the world if T did not waive, S would (weighted desire-types permitting) have admissible grounds to prefer that T should waive even when she and T preferred the same options in } E^{<S,U,Δ_{<S,T,U}>} \text{ and } E^{<T,U,Δ_{<S,T,U}>}. \text{ But this is not the case (at least in the God example). Remember that the consequences of T’s waiving, with the exception of those specified by the N-premises and N-conclusion, are inadmissible N-grounds for S.}\]
are deductively inferred from other premises. P1, P8, and P12 are premises in which S must have some credence given that she knows the obligations of the other members of N, holds them to be somewhat trustworthy, and hence holds them to be motivated to some degree to live up to their obligations.

P2, P9, and P6 are supported by psychological platitudes. P6 is also supported partly by an assumption that the members of N are rational, which I take to be implicit in the assumption that they are trustworthy. P6 is supported by a platitude about belief, which is that a rational person will have a positive credence in any contingent proposition. P2 and P9 are supported—albeit very weakly—by a platitude about motivation, which is that if someone can do something, she is more likely to do it if she has some desire to than if she has none.

P4 and P13 are entailed by the standardized terms of the relationships in N. The support that S gets for believing P7 also depends on S’s knowledge of these terms, and her ‘trusting’ beliefs that the other members of N know these rules and are rational. These admissible grounds licence S to infer, from the proposition that pro-φ entrusting party ‘x’ has a positive credence that ‘y’ prefers that x should waive, that x herself would rationally infer that there is some z (i.e. y) whose trust she would strain by not waiving.

Thus far I have been talking only about how ‘harmony’ among S’s preferences might be guaranteed. But ‘harmony’ among T’s preferences might be guaranteed in the same way. The set of admissible ground-types for T in E(S,T,Δ(S,T,U,φ)) could be identical to the set of admissible ground-types for S in E(S,T,Δ(S,T,U,φ)). T would have a discretion-licencing preference not to waive because T would prefer that U not be exposed to pressure from S to not-φ, and T’s waiving would increase (minutely) the likelihood that this would happen. In fact, the same argument (P1-C) that explained S’s preference in E(S,T,Δ(S,T,U,φ)), would give T the admissible difference-making credence in the same N-conclusion that T needed to prefer not to waive in E(S,T,Δ(S,T,U,φ)), as long as T preferred that U should φ in E(T,U,Δ(S,T,U,φ)).

Given the problematic and oversimplifying assumptions that my explanation for the realization of ‘harmony’ has relied upon, this explanation cannot be regarded as a complete success. The oversimplifying assumptions could, I think, have been avoided had I used a more elaborate argument as my example, but I know of no way to avoid a commitment to infinitesimal credences and regularity. At any rate, I think that I have made a tolerably plausible case for the possibility of guaranteeing ‘harmony’, and at
least given readers an idea of the challenges my project faces so that they may better evaluate it.

One cost of my account of how ‘harmony’ would be guaranteed that I have not addressed yet is that it requires the terms of the trust relationships shared by the parties to be quite complex. It might seem that the degree of complexity they require is unrealistic. People are capable of having third-party disagreements with each other who do not seem to have spent a lot of time together figuring out which ground-types they need to count as admissible in order to ensure ‘harmony’. This isn’t too puzzling though, for the measures needed to ensure ‘harmony’ will almost always be the same. We can, after all, say roughly which ground-types need to be admissible for members of N for this purpose without paying much attention to the content of members’ weighted desires, or the contingent circumstances in which they find themselves.

We can thus explain the ease with which we form dual-order fully connected networks, and other third-party-disagreement-enabling networks, by postulating that we have a pre-formed template for terms of such networks in our heads that we tacitly and easily apply in the base conditions in which we find ourselves. Such a template would remove the need, when forming trust networks, for careful design to guarantee ‘harmony’, though it would not remove the need for negotiation and bargaining with other (would-be) members. The template may be customary and learned, or we may possess it as part of a special disposition to form, or facility for forming, entrusted concern relationships and trust networks, acquired through biological evolution. I think that an evolved facility must be part of the explanation for our possession of the template, for when we teach our children about trust and about how to get on with others, we don’t seem to be teaching them how to design dual-order trust networks with a view to preserving ‘harmony’.

If this is right, our possession of the template would not make sense if forming ‘harmony’-preserving dual-order networks were not fitness-enhancing. But I think the story I have already supplied about the advantages of ‘harmony’ goes some way toward explaining why it would be. In networks that are formed to serve the common desires W of a largish group, in which each member is an entrusted party in an entrusted concern relationship with each other member, and the set of each member’s weighted desire-types qua entrusted and entrusting party is W, it is to the advantage of the group that members’ voiced preferences should be examined for bad reasoning and foul play. And it is to the advantage of individual members to scrutinize the reasoning behind other
members’ voiced preferences if they are able to, as a means to changing their discretion-licencing preferences and hence their behaviour in desired ways, and as a way to win esteem.

A second-order network will give them a more extensive ability to scrutinize one another’s preferences, and hence will increase overall levels of scrutiny. But this additional scrutiny will only amount to an increase in scrutiny of first-order parties’ preferences to the extent that flaws in a second-order party’s voiceable preferences are correlated with flaws of the same kind in a corresponding set of her voiceable preferences qua first-order party. ‘Harmony’ virtually guarantees this—it virtually guarantees that, for all φ, all network members S, all second-order entrusted parties T, and all first-order entrusted parties U, there are flaws in S’s preferences that T should waive (or not waive) her demand that U should φ, if and only if there are flaws in her preference that U should not-φ (or φ). (S may be the same person as T.)

Thus, if membership in largish groups with some shared ends was a circumstance in which ancestors of modern humans commonly found themselves, it is not too far-fetched to suppose that a facility for forming trust relationships that would serve these ends—trust relationships of the complex sort that I have been describing—would have been fitness-enhancing.

Conclusion

In this long chapter I have tried, for certain features of the moral trust network, to explain why trust networks with these features might exist, and to show that they are not entirely foreign to our experience. These features of the moral trust network are as follows: each member is both an entrusted party and an entrusting party in an entrusted concern relationship with each other member; the weighted desire-types of all parties in all these relationships are identical; members of this network try to change other members’ behaviour chiefly by changing the reasoning behind their preferences rather than by inducing them via trust-based pressures to yield to their preferences; members will be concerned to engage with the reasoning behind preferences even of members who are not themselves going to perform relevant actions; and members can have non-cognitive disagreements with one another about what other members should do. I have described a kind of trust network with all of these properties, viz., a dual-order fully connected network. I have given what I take to be a plausible example of such a network (the Seymours’ familial trust network). And I have explained why we should
sometimes expect such networks to exist. Specifically, I argued that concerns to avoid
collision between relationships in common entrusted party networks would motivate
people to form uniform weighted desires networks; and, later, that concerns to promote
the satisfaction of common desires and facilitate scrutiny of preferences would motivate
people to form dual-order fully connected networks composed of many overlapping
uniform weighted desires networks.
PART II

MORAL PSYCHOLOGY AND THE MORAL TRUST NETWORK
Moral Judgement

In Part II I develop and defend fiduciary non-cognitivism, my novel non-cognitivist theory of moral psychology. The definitive feature of this theory is that it holds that moral judgements at least partly consist of the voiceable preferences of members of a network of trust relationships linking all moralists. In the chapters that follow this one I will make the best case I can for the existence of the moral trust network, and try to explain the apparent rule-like character of moral requirements, the nature of substantive moral reasoning, the changes in motivation produced by moral reasoning, and the phenomenon of moral uncertainty, among other things.

In this chapter I will present the beginnings of a fiduciary non-cognitivist account of moral judgement. It is limited in the following ways: it concerns only judgements of basic moral principle, as opposed to ‘derived’ moral judgements whose rationality straightforwardly depends on contingent beliefs; it concerns only judgements of the sort that are expressed by simple indicative ‘atomic’ sentences predicating wrongness or permissibility; and it merely concerns the non-cognitive attitudes involved in moral judgements, not any cognitive components they may have. The basic account is set out in Section 4.1. Section 4.2 explains why inconsistent judgements of basic moral principle put people in disagreement.

4.1 The Non-cognitive Attitudes Involved in Moral Judgements

4.1.1 Moralists’ Voiceable Preferences

The account of moral judgements that I present in this chapter will have to rely on some assumptions about moralists’ voiceable preferences that I don’t defend until later chapters. I will describe these here, and introduce some related terminology.
The moral trust network is a dual-order fully connected network. In 3.2.1 I argued that one of the typical conditions for the formation of such a network was that prospective members should have certain desires in common, and that these are typically the desires that would count as weighted for network members. The relevant common desires of moralists are altruistic desires. These are desire, for instance, that people should not harm, deceive, make others suffer, treat others unfairly, and the like. Moralists’ weighted desires consist chiefly of these altruistic desires.

Unusually, for all I’ve said so far, the moral trust network does not simply give first-order members voices in one another’s choices of which actions to perform. Rather, it gives them voices in one another’s choices of which rules or ‘norms’ to conform to. I’ll say more about norms and what it means to conform to or follow them in Chapter 6. For now the following will suffice. A norm is a set of prescriptions either forbidding or requiring all actions with a certain set of properties. The extension of a norm is a set of ‘action-circumstance pairs.’ An action-circumstance pair is a pair of an action and a possible circumstance in which it is performed. An action-circumstance pair is ‘maximally specified’—it is compatible with exactly one possible world. Action-circumstance pair \(\langle \varphi_1, C_1 \rangle\) and action-circumstance pair \(\langle \varphi_2, C_2 \rangle\) are numerically identical just in case every detail about the actions, and about the circumstances in which they occur, is exactly similar. For all action-circumstance pairs \(\langle \varphi, C \rangle\) and all norms \(N, \langle \varphi, C \rangle\) is in the extension of \(N\) just in case \(N\) either forbids or requires \(\varphi\)-ing in \(C\).

An agent follows a norm if and only if she performs no action that it does not require of her.\(^{116}\) Thus she fails to follow the norm if she either performs an action it forbids, or because she acts in a circumstance that falls outside of its extension. This means that a norm that only forbids actions cannot be followed. Readers may treat ‘follows’ and ‘conforms to’ as interchangeable for present purposes; in Chapter 6 I will make my usage of ‘conforms to’ more specific, and this precisification will apply retrospectively to the present chapter.

Norms \(N\) and \(O\) make incompatible prescriptions in some circumstance \(C\) just in case either \(N\) requires an action in \(C\) that \(O\) forbids or \(N\) forbids an action in \(C\) that \(O\)

---

\(^{116}\) This is a technical usage. It is meant to avoid an ambiguity in the ordinary English phrase ‘follow a norm’. The English phrase could mean ‘do nothing that the norm does not require/permit’ (e.g. it makes sense to say, ‘How can I follow traffic rules when I’m not even on the road?’), or ‘do nothing that the norm forbids’ (e.g. it makes sense to say ‘I haven’t failed to follow traffic rules—how could I have when I wasn’t even on the road?’).
requires (otherwise they make compatible prescriptions in C). N is incompatible with O just in case there is some circumstance in which they make incompatible prescriptions.

The moral trust network is a dual-order fully connected network, in which weighted desire-types for all parties in all roles are the same and ‘harmony’ is preserved. This means that, for all moralists S, all moralists T, and all moralists U, such that S≠T≠U, and S is rational, among the following statements, (1) and (2) are equivalent, and (1) and (2) each entail (3) and (4):

1. S has a discretion-licencing preference qua first-order moral entrusted party that S should conform to N when in circumstances C.
2. S has an accommodation-obliging preference qua first-order moral entrusting party that U should conform to N when in C.
3. If U has a discretion-licencing preference to conform to N when in C and U has not yielded, S has a discretion-licencing preference qua second-order moral entrusted party that S should not waive her demand that U should conform to N when in C.
4. If U has a discretion-licencing preference not to conform to N when in C and has not yielded, and T has an accommodation-obliging preference that U should not conform to N when in C and has not waived, S has an accommodation-obliging preference qua second-order moral entrusting party that T should waive her demand that U should conform to N when in C.

Note that a preference for conforming to N when in C is a conditional preference. I discuss conditional voiceable preferences, and their capacity to put people in disagreement, in Section 4.2 below.

I will sometimes say that a moralist prefers, or has a preference for, a norm, as a shorthand for saying that she has a preference for conforming to that norm. I will call the set of norms N such that a moralist S has a voiceable preference that moralists should conform to them whenever such conformity is possible, S’s norms.

4.1.2 How Much I Need to Say about Moral Judgements

I will not be giving a complete account of moral judgement, because there are some properties of moral judgements about which I wish to remain agnostic. What account a non-cognitivist gives of moral judgements will affect her theory of the semantics of moral statements—it may affect her capacity to solve the Frege-Geach problem, for instance. I don’t know what sort of account of moral judgement a non-cognitivist must give if she is to provide an adequate moral semantics. So I don’t want to say more about the nature of moral judgement than I have to, given my more limited aims of explaining moral disagreement in matters of basic principle, certain types of moral reasoning, and certain features of our moral practices. Specifically, I don’t want to take a stand on the question of whether moral judgements consist merely of non-cognitive attitudes, or
whether (as I suspect) they consist of combinations of cognitive and non-cognitive attitudes. And I don’t need to take a stand on this issue given my aims.

Although I shall not be giving a theory of the semantics of moral sentences or of deductive moral reasoning, I do need to ensure, as best I can, that what I say about the nature of moral judgement is compatible with some adequate theory of both of these subject matters. This means, among other things, that I don’t want to give an account of the non-cognitive attitudes that are involved in judgements of permissibility that would leave non-cognitivists unable to solve the negation problem.\footnote{Unwin, ‘Norms and Negation’; Schroeder, Being For, 44-9.} This is the problem of explaining why it is rationally incoherent to simultaneously judge that φ-ing is wrong and that φ-ing is permissible, and hence why it is inconsistent to say that φ-ing is wrong and that φ-ing is permissible, while preserving a distinction between judging that φ-ing is permissible and being unopinionated about the morality of φ-ing.

I also have a reason to be concerned with this problem that falls squarely within the scope of the present work: unless, on my account, it is rationally incoherent to judge that φ-ing is wrong and that φ-ing is permissible, where these are judgements of basic principle, my theory of non-cognitive disagreement will not predict that two such judgements, when held by different people, put those people in disagreement. This is because (i) I presume that, for the non-cognitivist, disagreement in matters of basic moral principle necessarily involves robust disagreement in preferences rather than mere cognitive disagreement; (ii) my theory of robust non-cognitive disagreement says that people disagree in preference robustly only if they have mutually unsatisfiable preferences; and (iii) when mutually unsatisfiable preferences are held by the same person, that person will be rationally incoherent. So I need to ensure that intuitively inconsistent pairs of permissibility and wrongness judgements, held by the same person, are genuinely incoherent, if I am to explain disagreement in basic principle.

I shall in large part adopt a solution to the negation problem proposed by Jeremy Schwartz and Christopher Hom.\footnote{Jeremy Schwartz and Christopher Hom, ‘Why the Negation Problem is Not a Problem for Expressivism’, Nous (2015): 824-45.} The following summary of their solution is rough, and also skewed in a particular direction, for I am treating it as a solution to the negation problem understood as a problem for non-cognitivist moral psychology, whereas the authors are focused on the negation problem as a problem for expressivist semantics. I
thus disregard some features of the proposal which according to the authors make it semantically virtuous.

4.1.3 The Negation Problem, and ‘Arrangements of Desires’

Schwartz and Hom propose that non-cognitivists should construe judgements of moral requirement, permissibility, and impermissibility as attitudes of the same type held toward different objects. A judgement that \(\phi\)-ing is required is some sort of pro-attitude (call it a ‘pro-attitude\(^*\)) toward \(\phi\)-ing under all ‘arrangements of desires’\(.)^{119}\) Intuitively, this can be interpreted as \(\phi\)-ing whether one wants to or not. A judgement that \(\phi\)-ing is permitted is a pro-attitude\(^*\) toward \(\phi\)-ing under some arrangements of desires.\(^{120}\) The authors sometimes suggest \(\phi\)-ing if one wants to as a more specific account of the object of the pro-attitude\(^*\) involved in judging that \(\phi\) is permissible.\(^{121}\) A judgement that \(\phi\)-ing is wrong is a pro-attitude\(^*\) toward \(\phi\)-ing under no arrangement of desires.\(^{122}\) (To judge that \(\phi\)-ing is merely permissible, i.e. permissible and not required, is presumably to have pro-attitudes\(^*\) both toward \(\phi\)-ing if one wants to and toward not-\(\phi\)-ing if one wants to not-\(\phi\).)

How do Schwartz and Hom propose to solve the negation problem? That is, how do they give an account of (1) judging that \(\phi\) is permitted, (2) judging that \(\phi\) is wrong, and (3) being undecided about the permissibility of \(\phi\), that predicts that an agent who does both (1) and (2) at the same time is irrational, and that preserves a distinction between (1) and (3)? This requires that pro-attitudes\(^*\) be attitudes that a person cannot rationally hold toward both \(\phi\)-ing and not-\(\phi\)-ing. Suppose that this desideratum is met. It is then irrational to simultaneously judge that \(\phi\)-ing is permissible and impermissible because it is irrational to simultaneously have a pro-attitude\(^*\) toward \(\phi\)-ing under some arrangement of desires and a pro-attitude\(^*\) toward \(\phi\)-ing under no arrangement of desires. To be undecided about the propriety of \(\phi\)-ing is to lack both a pro-attitude\(^*\) toward \(\phi\)-ing under some arrangement of desires and a pro-attitude\(^*\) toward \(\phi\)-ing under no arrangement of desires. This mental state is clearly different from the state of judging that \(\phi\)-ing is permissible on their account.

---

119 Throughout their discussion Schwartz and Hom treat what I am calling ‘pro-attitudes\(^*\)’ as attitudes of approval, but they do not commit to an account of their nature (‘Why the Negation Problem’, 835). It is not important that the attitude type that figures in the theory should be a pro-attitude; a con-attitude would work just as well mutatis mutandis.

120 Schwartz and Hom, ‘Why the Negation Problem’, 832.


I shall posit that the non-cognitive attitudes that are involved in all moral judgements, and that explain the irrationality of holding basic moral principles that are, intuitively, inconsistent, are voiceable preferences of members to the moral trust network as such. They are in fact combinations of voiceable preferences held by moralists in each of their roles qua parties to the network, i.e. as first- and second-order moral entrusted and entrusting parties. But I won’t worry about moralists’ preferences qua parties to the second-order network for the moment. And I needn’t distinguish between a moralist’s preferences qua entrusted and entrusting party. These ought both to be voiceable preferences for conformity to the same norms, so I can talk simply of ‘moralists’ preferences’, without specifying whether they are accommodation-obliging or discretion-licensing.

As I have indicated, to solve the negation problem in the way that Schwartz and Hom propose, a non-cognitivist theory must characterize pro-attitudes* in such a way that they meet the following desideratum: that one cannot coherently hold a pro-attitude* toward both φ-ing and not-φ-ing. Voiceable preferences meet this desideratum: one cannot coherently voiceably prefer conformity to norms requiring φ and norms forbidding φ.

I must also give a characterization of the ‘arrangements of desires’ that, if I avail myself of Schwartz and Hom’s proposal, figure in the content of, and serve to distinguish, the non-cognitive attitudes that are involved in judgements of requirement and permissibility. As a first approximation, I propose that ‘arrangements of desires’ are circumstances of having some pre-trust preferences rather than others. Thus a judgement that B is required to φ is, roughly, a voiceable preference that B should follow norms requiring φ-ing whether one pre-trust prefers to φ or not. A judgement that it is permissible for B to φ is, roughly, a voiceable preference that B should follow norms requiring φ-ing if one pre-trust prefers to φ. A judgement that it is wrong for B to φ is, roughly, a voiceable preference that B should follow norms requiring not φ-ing whether one pre-trust prefers to φ or not.

Why not characterize ‘arrangements of desires’ as states of having some all-things-considered preferences rather than others? There is a problem with this view. Suppose that a moral entrusted party B has not yet formed a discretion-licencing preference between norms forbidding and norms requiring φ. He has an all-things-considered preference to φ. But B is, as I shall say, scrupulous: he is a moralist who complies with the terms of his trust relationships qua member of the moral trust
network. He thus does not exceed his discretion qua moral entrusted party. This means that he would lose his all-things-considered preference to φ if he formed a discretion-licencing preference to follow norms forbidding φ. He has a conditional all-things-considered preference to not-φ if and only if he has (i) a trust-based obligation to not-φ and (ii) no trust-based obligation to φ. He only has an all-things-considered preference to φ because (ii) is false: the current state of his voiceable preferences is such that to avoid exceeding his discretion he must, impossibly, both φ and not-φ. Given that he cannot avoid violating trust, he prefers, for non-trust-related reasons, to φ.

Suppose that B comes to judge that φ-ing is merely permissible—he comes to form a voiceable preference for norms requiring him to φ iff he all-things-considered prefers to φ. Do his preferred norms require him to φ? It might seem obvious that they do. B had an all-things-considered preference to φ before he formed his judgement that φ is permissible, and nothing seems to have changed that would alter this preference. However, what B has an all-things-considered preference to do determines what he has a trust-based obligation to do, and (since B is a scrupulous moralist) whether or not B has an all-things-considered preference to φ depends on whether or not he has a trust-based obligation to φ. So actually, we can’t establish that B would have an all-things-considered preference to φ in this situation without presupposing that B has an all-things-considered preference to φ. The view that ‘arrangements of desires’ are all-things-considered preference states thus seemingly makes scrupulous moralists’ obligations indeterminate in some sense. The hypothesis that ‘arrangements of desires’ are pre-trust preference states avoids this sort of muddle.

However, this hypothesis seems to be defective for a different reason, at least given my other commitments. I am committed to the idea that preferences can be indeterminate—my explanation for the phenomenon of moral indecision depends on this (see 6.4.1 and Chapter 8 below)—and so I am committed to the possibility that pre-trust preferences can be indeterminate. If ‘arrangements of desires’ are pre-trust preference states, it is not clear whether my account of the non-cognitive attitudes involved in permission, requirement, and prohibition judgements will be compatible with any intuitively satisfactory account of those judgements insofar as they pertain to agents with indeterminate pre-trust preferences.

For instance, suppose that Oliver faces a choice between stealing to relieve extreme hunger or refraining from doing so, and that his pre-trust preferences do not discriminate between these options due to indeterminacy. Further, suppose that
(a) the non-cognitive attitude involved in a judgement that stealing to relieve extreme hunger is permissible is a voiceable preference for conformity to norms requiring an agent to steal to relieve extreme hunger if she pre-trust prefers to; and

(b) the non-cognitive attitude involved in a judgement that not stealing to relieve extreme hunger is permissible is a voiceable preference for conformity to norms requiring an agent to refrain from stealing to relieve extreme hunger if she pre-trust prefers to so refrain.

Here are two options for characterizing the attitudes involved in the judgement that stealing to relieve extreme hunger is wrong:

(c) the non-cognitive attitude involved in a judgement that stealing to relieve extreme hunger is wrong is a voiceable preference for conformity to norms forbidding an agent to steal to relieve extreme hunger if she either pre-trust prefers to steal to relieve extreme hunger or pre-trust prefers not to steal to relieve extreme hunger.

(c′) the non-cognitive attitude involved in a judgement that stealing to relieve extreme hunger is wrong is a voiceable preference for conformity to norms forbidding an agent to steal to relieve extreme hunger if she either pre-trust prefers to steal to relieve extreme hunger or does not pre-trust prefer to steal to relieve extreme hunger.

Observe that if (c) is true, and Oliver judges that stealing to relieve extreme hunger is wrong, he would take this judgement to be irrelevant to his decision. This is because the non-cognitive attitude involved in this judgement is a preference for following norms forbidding agents with determinate pre-trust preferences from stealing, and he lacks such preferences. But intuitively his wrongness judgement should be relevant to his decision.

Now suppose instead that (c′) is true, and that Oliver judges that stealing to relieve extreme hunger is merely permissible—that is, he judges that both stealing and not stealing to relieve extreme hunger are permissible. The non-cognitive attitudes involved in this judgement are the voiceable preferences described in (a) and (b). Oliver will thus take his permissibility judgement to be irrelevant to his choice; he is not in the circumstances these voiceable preferences apply to. Indeed, he will be forced to violate trust, either by resisting the accommodation-obliging preferences of those who judge that stealing in his situation is required, or by resisting the accommodation-obliging preferences of those who judge that stealing in his situation is wrong, without a discretion-licencing preference to do so. This also seems unsatisfactory. Intuitively, he ought to be able avow his permission judgement in defence of whichever choice he makes when other moralists raise objections. But this would only make sense, according to fiduciary non-cognitivism, if by avowing this judgement he were voicing a
discretion-licencing preference for his chosen option, and on our present assumptions he would not be doing so.

There might be, or might seem to be, ways to avoid these problems on the pre-trust preferences account, but at best they create similarly bad problems of their own. For instance, one might take the non-cognitive attitude involved in judging that φ-ing is permissible to be a preference for following norms requiring you to φ if you don’t have pre-trust preferences not to φ. This would avoid making permissibility judgements inapplicable to people with indeterminate pre-trust preferences. But it would also make rational judgements of mere permissibility impossible. On this view (a) and (b) would need to be revised thus:

(a’) the non-cognitive attitude involved in a judgement that stealing to relieve extreme hunger is permissible is a voiceable preference for conformity to norms requiring an agent to steal to relieve extreme hunger if she does not pre-trust prefer not to steal.

(b’) the non-cognitive attitude involved in a judgement that not stealing to relieve extreme hunger is permissible is a voiceable preference for conformity to norms requiring an agent to refrain from stealing to relieve extreme hunger if she does not pre-trust prefer to steal.

To judge that stealing to relieve extreme hunger is merely permissible is to hold both of these preferences. But they are rationally inconsistent: they are preferences for following norms that make conflicting prescriptions in Oliver’s case, for instance.

The problem with taking ‘arrangements of desires’ to be arrangements of pre-trust preferences among available options is that it is possible to lack such preferences. ‘Arrangements of desires’ must be unavoidable features of any agent’s circumstances. I shall instead characterize ‘arrangements of desires’ as one’s psychological dispositions to act or behave before one’s moral judgements are taken into account. Even if an agent’s preferences are indeterminate, she must still do something, even if that thing is dithering or standing in a state of rational paralysis while torn between options. It may, perhaps, also be possible for an agent whose preferences are indeterminate, or equally balanced between two options, to voluntarily and non-rationally ‘plump’ for one option rather than another. I take it that, if these processes and behaviours—these agential ‘doings’—occur, they occur as a result of psychological dispositions of some sort. (Such dispositions to behave and plump seem to differ from desires in at least the following respect: they are not capable of combining with beliefs in rational ways to form other dispositions.)
So there seems to be a broad class of ‘agential dispositions’ (as I shall call them) that includes but is not limited to preferences, and which have the desirable property that pre-trust preferences lack: necessarily, if it is up to an agent whether she φs or not-φs at t, that agent has a (determinate) agential disposition to φ or not φ at t. Let me relabel the things I have been calling ‘arrangements of desires’ as ‘dispositional circumstances’, so that it will not seem odd for me to posit that they can differ with respect to changes in non-conative agential dispositions. The agential dispositions that dispositional circumstances consist of on my theory are what I will call pre-trust agential dispositions. These are the agential dispositions a moralist has ‘before’ her trust-based desires qua moralist are taken into account.

Now I must make a postulation about the way we ordinarily talk about circumstances when we describe or state the content of moral judgements. I must say that when in ordinary talk we say that φ-ing in a circumstance C is permissible, and C is what we prereflectively think of as a specific circumstance, C is typically not really a specific circumstance. It will be specific in most respects, but typically it will underdetermine the agent’s dispositional circumstances. More precisely, C will typically be a set of possible circumstances that are all exactly similar, except as regards what the agent is pre-trust disposed to do.

I will call circumstances like C, which we prereflectively think of as specific, but which on my theory do not determine the agent’s pre-trust dispositions, dispositionally non-specific circumstances. I posit that normally when we make moral requirement and impermissibility judgements about actions in what we take to be specific circumstances, the circumstances in question are dispositionally non-specific (except insofar as the explicitly specified circumstances ‘rule out’ certain dispositionally specific circumstances—that is, except insofar as the proposition that the explicit circumstances obtain entails that the dispositionally specific circumstances do not obtain). And I posit that when we make more general moral judgements about φ-ing in explicitly non-specific circumstances (e.g. when I judge that lying when doing so will save a life is always wrong, rather than wrong in a particular case), the circumstances in question can normally be construed as consisting exclusively of sets of dispositionally non-specific circumstances (e.g. the set of all dispositionally non-specific circumstances in which lying will save a life). So the more general circumstances that figure in statements of moral principles are normally—to use my jargon elastically—dispositionally non-specific too. Dispositionally non-specific circumstances differ from the sorts of
circumstances that figure in the action-circumstance pairs that make up the extensions of norm-partitions. The latter circumstances do specify the agent’s pre-trust dispositions—they are ‘dispositionally specific’.

I will now give my more thorough account of the non-cognitive attitudes that are involved in judgements of moral requirement, permissibility, and impermissibility. The account applies only to judgements of basic principle. This is indicated by the specification that they are judgements about what is necessarily required, permitted, and wrong. The account also applies only to judgements about actions in ‘specific’ (but perhaps dispositionally non-specific) circumstances. In what follows, ‘C’ is a ‘specific’ circumstance of this sort. All ‘parties’ are, of course, parties to moral entrusted concern relationships. A set of circumstances $C_1$ entails a set of circumstances $C_2$ just in case the proposition that $C_1$ obtains implies that $C_2$ obtains (i.e. $C_1$ is a subset of $C_2$).

**Requirement.** S judges that $\varphi$-ing in C is necessarily required only if, for all $C^*$ such that $C^*$ is a dispositionally specific circumstance that entails C, S has

(i) a voiceable preference qua first-order entrusted party to follow norms requiring $\varphi$-ing in $C^*$,

(ii) a voiceable preference qua first-order entrusting party that first-order entrusted parties should follow norms requiring $\varphi$-ing in $C^*$,

(iii) a voiceable preference qua second-order entrusted party not to waive her entitlement to demand that first-order entrusted parties should accommodate her preference that they should follow norms requiring $\varphi$-ing in $C^*$, and

(iv) a voiceable preference qua second-order entrusting party that second-order entrusted parties who have voiceable preferences qua first-order entrusting parties that first-order entrusted parties should follow norms forbidding $\varphi$-ing in $C^*$ should waive their entitlements that first-order entrusted parties should accommodate these preferences.

**Permissibility.** S judges that $\varphi$-ing in C is necessarily permissible only if, for all $C^*$ such that $C^*$ is a dispositionally specific circumstance that entails C, and $C^*$ specifies that the agent is pre-trust disposed to $\varphi$, S has

(i) a voiceable preference qua first-order entrusted party to follow norms requiring $\varphi$-ing in $C^*$,

(ii) a voiceable preference qua first-order entrusting party that entrusted parties should follow norms requiring $\varphi$-ing in $C^*$,

(iii) a voiceable preference qua second-order entrusted party not to waive her entitlement to demand that first-order entrusted parties should accommodate her preference that they should follow norms requiring $\varphi$-ing in $C^*$, and

(iv) a voiceable preference qua second-order entrusting party that second-order entrusted parties who have voiceable preferences qua first-order entrusting parties that first-order entrusted parties should follow norms forbidding $\varphi$-ing in $C^*$ should waive their entitlements that first-order entrusted parties should accommodate these preferences.

**Impermissibility.** S judges that $\varphi$-ing in C is necessarily wrong only if, for all $C^*$ such that $C^*$ is a dispositionally specific circumstance that entails C, S has
(i) a voiceable preference qua first-order entrusted party to follow norms forbidding φ-ing in C*,
(ii) a voiceable preference qua first-order entrusting party that entrusted parties should follow norms forbidding φ-ing in C*,
(iii) a voiceable preference qua second-order entrusted party not to waive her entitlement to demand that first-order entrusted parties should accommodate her preference that they should follow norms forbidding φ-ing in C*, and
(iv) a voiceable preference qua second-order entrusting party that second-order entrusted parties who have voiceable preferences qua first-order entrusting parties that first-order entrusted parties should follow norms requiring φ-ing in C* should waive their entitlements that first-order entrusted parties should accommodate these preferences.

4.1.4 The Presumption of Dispositional Non-specificity

According to fiduciary non-cognitivism, moral judgements about actions in what we normally take to be ‘specific’ circumstances are, in part, and presumptively, preferences for conformity to sets of norms making prescriptions about what to do in circumstances that do not fully specify the agent’s pre-trust dispositions, but which are maximally specific in all other respects. I now wish to say something about the ‘presumptively’ qualifier. This is supposed to signify that it is not a necessary feature of a moral judgement about an action in a ‘specific’ circumstance that these circumstances are dispositionally non-specific. The ‘specific’ circumstances to which such moral judgements apply may actually be as specific as the circumstances to which moralists’ preferred norms apply—that is, they may fully specify the agent’s pre-trust dispositions. If they concern what to do in circumstances of this sort, judgements permitting and requiring the same action will be identical and indistinguishable. Alternatively the relevant ‘specific circumstances’ may not be fully specific as regards pre-trust dispositions, but may still be more specific in this regard than is conventionally presumed. They may narrow down the range of dispositional circumstances of the actions they apply to, instead of being entirely silent about them.

Fiduciary non-cognitivism needs to allow this if it is to account for the diversity of our moral judgements. The prescriptions that preference utilitarians give an agent could potentially vary depending on that agent’s own preferences. On some views, whether someone is permitted or required to torture a terrorist for information about a ticking bomb might depend on whether or not the agent desires to torture the terrorist—having such a desire might make the action wrong. An accurate characterization of the dispositions that are part of the circumstances of the actions to which these judgements apply would presumably not say anything about pre-trust dispositions specifically, but it
would presumably narrow down the possible dispositional circumstances that the parties can occupy. The torture example points toward an odd possibility, namely, that a moralist might judge that a moralist S is required to φ just in case S is pre-trust disposed to not-φ. On my account, this is not a special kind of judgement (e.g. an ‘anti-permission’ judgement); it is merely a set of requirement judgements applying to dispositional circumstances that are more fully specified than usual.

We could exhaustively describe any moralist’s voiceable preferences by talking only of the requirement and impermissibility judgements she accepts, where those judgements pertain to genuinely fully specified circumstances. Permissibility judgements are thus reducible to requirement judgements applying to genuinely fully specified circumstances. Our conceptual schemes make room for permissibility judgements as a salient category of moral judgement because this facilitates more efficient moral discourse (and perhaps moral thinking as well). Most of the time, the efficient reporting and voicing of voiceable preferences is made easier if (a) moralists all presume that the circumstances to which moral judgements explicitly apply are dispositionally non-specific, and (b) when more specific information about the dispositional circumstances to which moral judgements apply needs to be communicated, this is normally done, not by elaborating on the explicit circumstances, but by specifying that the judgement is of a certain kind, viz., a permission judgement.

Why is the information about dispositional circumstances that is communicated by specifying that a moral judgement is a permission judgement of such special interest that it warrants being communicated in this peculiar way? Relatedly, why is it relatively unusual for moralists to want to communicate any sorts of information about the dispositional circumstances of the actions to which their voiceable preferences apply, unless it is the information that they communicate by talking of permissions and permission judgements? I cannot provide an independently plausible answer these questions; all I can do is postulate and hypothesize. Accordingly, I hypothesize that the answer lies in the content of our altruistic desires. The extent to which an agent S’s φ-ing satisfies our altruistic desires does not, ordinarily and in most respects, depend on facts about the agential dispositions S has when she φs. But there is one possible feature of S’s dispositional circumstances that would be fairly likely to affect the extent to which her φ-ing satisfies our altruistic desires, viz., her having a pre-trust disposition to φ. S’s φ-ing is less likely to frustrate our altruistic desires if she does it in accordance with a pre-trust disposition than if she does it out of a desire not to violate trust. Perhaps
our altruistic desires are generally frustrated to some degree by instances in which trust-based obligations induce people to do otherwise than they would be disposed to do if left alone—that is, by trust-based impositions. Unfortunately this somewhat intuitively appealing suggestion would, if true, create complications that I cannot address at present, so I must leave it undeveloped.

To reiterate, I have hypothesized that altruistic desires are in many cases less well satisfied by deeds done contrary to pre-trust dispositions, and largely unaffected by the dispositional circumstances of actions in other respects. This hypothesized fact does not necessarily mean that moralists will prefer conformity to norms that disproportionately require people to perform actions that they are pre-trust disposed to perform, but whose prescriptions do not depend on pre-trust dispositions in other respects. This will only be a tendency; and considerations of simplicity will sometimes favour norms that defy this tendency. Certain altruistic desires may militate against this tendency too—for instance, the desire that people shouldn’t indulge their (pre-trust) desires to torture. But the tendency, if strong enough, would explain the special place of permission judgements among moral judgements.

4.2 Disagreement in Matters of Basic Principle

Moralists as such can disagree about what people ought to do in purely hypothetical situations, and about what people should have done in the past. Fiduciary non-cognitivism must offer an explanation for these sorts of disagreement if it is to be plausible, but this is beyond the scope of the present work. I only aspire to explain more obviously practical disagreements, those that have some ‘direct’ relevance to deliberation and planning. I hope nonetheless to predict the truth of the following claim, which I shall call Necessary Rational Disagreement: rational people with conflicting fundamental moral principles necessarily disagree. This means predicting that they disagree even when no probable choice situation is in the offing about which their principles give conflicting prescriptions. In such cases, as I shall argue, their principles must nonetheless put them in robust non-cognitive disagreement about what to do in situations that are highly improbable.

In order to predict this, I need the following thesis to be true: that rational people with opposing voiceable preferences concerning a choice situation that they believe is extremely unlikely to eventuate, are necessarily in (robust) non-cognitive disagreement, these beliefs notwithstanding. I cannot think of natural-seeming cases with which to
thoroughly test this thesis—that is, cases involving fantastically small positive probabilities. Nonetheless, I can at least test it against cases in which people disagree about what to do in circumstances that they find improbable. Consider a variation on Stevenson’s invitation case. As before, Mr. and Mrs. A are deliberating about whom to invite to their party.

*Invitation (Conditional Disagreement)*

*Mrs. A:* It’s possible that Mr. Toff will be back from India tomorrow.

*Mr. A:* That’s very unlikely.

*Mrs. A:* I know. But if he does get back, we could invite him.

*Mr. A:* I’d still rather we invited Mr. Chum. He’s an old friend.

*Mrs. A:* But as you know, Mr. Toff’s connections are excellent. I’d prefer that we invited Toff.

*Mr. A:* We disagree then.

*Mrs. A:* I don’t think we disagree. We just prefer different things.

Like the original version, this seems to be a case of robust non-cognitive disagreement. So mutually unsatisfiable voiceable preferences about what to do in highly improbable situations do seem to suffice for non-cognitive disagreement.

If the thesis I have just argued for is true, it must be the case that people can be in robust non-cognitive disagreement even if the anticipated choice situation about which they disagree never actually eventuates. Suppose that Mr. Toff will not be back from India tomorrow, but, being unaware of this, Mr. and Mrs. A nonetheless have the voiceable preferences reported in *Invitation (Conditional Disagreement)*. I need it to be the case that these preferences put them in disagreement, despite the fact that they won’t have the option of inviting Mr. Toff anyway. But this seems very plausible. Our intuition that Mr. and Mrs. A disagree in *Invitation (Conditional Disagreement)* doesn’t seem to depend on our assuming that Mr. Toff might return from India.

Now I shall spell out why it is that rational moralists with conflicting basic principles necessarily disagree. Suppose that S prefers that B should follow norms N_S, that T prefers that B should follow norms N_T, and that N_S and N_T make incompatible prescriptions in a circumstance C_I and compatible prescriptions in circumstance C_C. (B might be the same person as S or T.) Observe that S’s and T’s preferences are not mutually unsatisfiable, because if B is in C_C he can accommodate both of them—he can follow both N_S and N_T. Since it is always possible that B will be in C_C, and since S and T might be certain that he will always be in C_C, S and T do not necessarily disagree about what B should do at any point in time. Indeed, for all I’ve said, and holding fixed
their voiceable preferences, they don’t necessarily disagree about anything. However, if S and T have some positive credence that T will find himself in C_t, the voiceable preferences of S and T will put them in disagreement about what B should do in C_t, at least if they have proportioned their preferences to their credences as rationality requires. This is an implication of the thesis that opposing voiceable preferences put people in disagreement even if they take the choice situation they disagree about to be extremely improbable.

S and T rationally must have some positive credence that B will be in C_t, because this is not an impossible state of affairs. Indeed, they rationally must have some positive credence that, for all C such that C is a circumstance in which N_S and N_T give incompatible prescriptions, and all B* such that B* is a moral entrusted party, and all future times t, B* will find himself in C at t. And they will disagree about whether any given B* should follow N_S or N_T in C at t. So every point of incompatibility between their norms must be a point of disagreement between them.

This is one way in which fiduciary non-cognitivism predicts that rational people with conflicting fundamental moral principles necessarily disagree. As I noted above, fiduciary non-cognitivism must, if it is to be plausible, also predict that moralists with conflicting fundamental principles disagree about what should be done in merely counterfactual and ‘past-tense’ choice situations. Such an explanation will imply Necessary Rational Disagreement, and hence might seem to make the argument I gave above as to how fiduciary non-cognitive predicts Necessary Rational Disagreement redundant. But the argument may still play an important role in explaining this datum. It may reduce the work that the fiduciary non-cognitivist has to do to explain moral disagreement about counterfactual actions. Since it is hard to see how anything could be a counterfactual moral judgement unless it were capable of disagreeing with other things of the same kind, the argument may reduce the work of explaining why moral thought and talk about counterfactual actions exists at all.

Conclusion

In this chapter I have postulated that the non-cognitive attitudes involved in moral judgements are voiceable preferences. Among judgements of basic principle, I have distinguished permissibility judgements from requirement judgements by positing that, in dispositionally non-specific circumstances, the former involve voiceable preferences for performing an action whenever the agent has a pre-trust disposition to do so, while
that latter involve voiceable preferences for performing an action regardless of the agent’s pre-trust dispositions. I have proposed that the salient category of permission judgements exists because our altruistic desires have a peculiar tendency to be frustrated, or less well satisfied, when certain actions are done in defiance of pre-trust dispositions. And I have explained why rational people whose judgements of basic moral principle are incompatible necessarily disagree.
5

Moral Universality

In this chapter I try to account for the apparent universality of morality: the apparent fact that anyone—any sane and intelligent human of a sufficient age at least—can sensibly make judgements about how any similar person morally ought to behave.\(^{123}\) In the context of my theory, this means defending my postulation of a moral trust network linking everyone who makes moral judgements. This postulation faces two main challenges. The moral trust network is of an extraordinary size. It links people who are and will always be complete strangers to one another. It links members of populations who may never have been in contact with one another even indirectly—for instance, it links moralists on Earth to extra-terrestrial moralists, if such exist. The first challenge, and by far the more pressing, is to explain how any such institution could exist—how it could have come about, and why it would persist and acquire new members.

The moral trust network has some features that are not shared by the fully connected networks that I discussed in Chapter 3. This is a cost if I cannot provide evidence that there are other trust networks with similar properties. I am plausibly committed to the empirical claim that there are constraints on the forms that trust networks are likely to take. This is because I have posited (3.3.3) that humans’ dispositions to form trust networks are partly explained by their possession of a special

\(^{123}\) Another proposition which we might want to label ‘moral universality’ is that a moral judgement about how one person ought to behave is implicitly a judgement about how anyone else ought to behave in the same circumstances. (Cf. Hare on the ‘covert universality’ of ought sentences. Hare, *The Language of Morals*, 154.) If I can fully account for moral universality of the sort mentioned in my opening sentence, I should be able to account for moral universality of this latter sort, provided that (i) ‘anyone’ is not understood *too* inclusively, (ii) my unfinished account of moral judgements is completed in a sensible way, and (iii) my argument in Chapter 6, especially 6.3.2, that the moral trust network would satisfy the condition which I call ‘Invariance’ is successful.
evolved facility for forming them. The varieties of trust networks this facility might dispose us to form is probably limited (just as variety among natural languages is presumably constrained by the features of humans’ evolved facility for linguistic communication). Thus a second challenge for the theorist who posits a moral trust network is to show that its novel features have non-moral precedents, and so to avoid ad hoc implications concerning the nature of our evolved facility for forming trust networks.

I only partly meet these challenges in this chapter, and indeed in this dissertation as a whole, so my postulation of the moral trust network does make ad hoc predictions about the nature of our evolved facility for forming trust relationships and trust networks. However, I attempt to address the first challenge in Section 5.1. There I argue that the ‘base conditions’ for a dual-order fully connected network exist among moralists, and suggest a very brief and speculative evolutionary explanation for the advent and persistence of the moral trust network among humans. In 5.2 I partly address the second challenge, giving examples of what I contend (on the basis of admittedly meagre evidence) are non-moral dual-order fully connected trust networks whose members, like those of the moral trust network, promiscuously extend their trust and commitments to anyone who meets certain conditions, even if they are perfect strangers. Since my responses to the two challenges are not wholly compelling, the plausibility of my postulation of the moral trust network depends rather heavily on an independent argument for the existence of the network that I offer in 5.3. There I argue that we need to postulate the moral trust network, or something similarly extravagant, if we are to provide a satisfying explanation of the phenomena of moral approval and disapproval. In 5.4 I consider some objections to my explanation for moral universality.

5.1 Why Would the Moral Trust Network Exist?

In 3.2.1 I proposed that the salient base conditions for a fully connected network among a group G were these: (i) G has significant shared desires W, which vary in strength among members of G; and (ii) all members of G can significantly affect the extent to which W are satisfied. I also proposed that any fully connected network was likely to be a dual-order network. If the arguments I gave for these proposals were credible, then the plausibility of my proposal that there is a dual-order fully connected network linking all moralists will be improved if I can plausibly claim that the base conditions for fully connected networks obtain among moralists.
When I gave this account of the base conditions for fully connected networks in 3.2.1, I was, as I admitted then, assuming that the members of a fully connected network were all able to communicate with one another. This assumption does not hold for the moral trust network. The failure of this assumption must complicate any explanation I might give for the existence of the moral trust network. But before I grapple with these complications, let us falsely assume for the time being that all moralists are able to communicate with one another, and would have been able to communicate with each other prior to the advent of the moral trust network.

As I foreshadowed in the last chapter, I propose that the set of desire-types that are shared by the group M of potential moralists, and that largely explain their participation in the moral trust network are what we might broadly call altruistic desires. They desire, for instance, that sentient creatures should not suffer, that intelligent creatures that knowingly cause sentient creatures to suffer should not benefit from doing so, and that intelligent creatures should not deceive other intelligent creatures. Indeed, they have altruistic desires corresponding to all our ‘moral intuitions’.

Thus, for any given member of M, and for every \( \varphi \) such that (i) \( \varphi \) is describable in non-moral terms, and (ii) the claim ‘\( \varphi \)-ing is necessarily morally wrong’ is intuitively plausible to most people, the M-member desires that intelligent creatures should never \( \varphi \). According to fiduciary non-cognitivism, these altruistic desires are the things that philosophers call intuitions of moral wrongness. They are basic desires. They are also ‘universal’, by which I mean that a statement of the content of an altruistic desire will contain no proper nouns or indexicals, and hence no references to particular times, people, or places. M-members’ altruistic desires vary in strength from member to member just as moral intuitions vary from person to person.

Members of M also have altruistic desires corresponding to intuitions of moral permissibility, although the correspondence here may not be quite as straightforward as in the impermissibility case. These are also basic and universal. In the context of the moral trust network, they give rise to derived desires that people should not do otherwise than they are pre-trust disposed to do qua moral entrusted parties, in certain matters. These derived desires more closely correspond to intuitions of permissibility—on my account, they are the things that philosophers call intuitions of moral permissibility. I am not quite sure what to say about the content of the basic altruistic desires they are derived from; they may, as I suggested in 4.1.4, be desires that people should not be subject to trust-based impositions in certain circumstances; or they may
have nothing to do with trust in particular. (When I talk of ‘altruistic’ desires in what follows, I have in mind only basic desires, not any desires that are derived from them, unless I indicate otherwise.)

By simply postulating altruistic desires, I give an extremely superficial explanation of a phenomenon for which other metaethicists have felt obliged to give a deep explanation, namely, the appearance of human altruism.\textsuperscript{124} Whether the superficiality of my explanation for human altruism is a cost for my theory depends on whether or not biology supplies a deep explanation that is compatible with my superficial one—that is, whether my posited altruistic desires are well explained by evolutionary mechanisms like reciprocal altruism, kin selection, and group selection.\textsuperscript{125} It is worth pointing out though that fiduciary non-cognitivism does provide a deep explanation for a \textit{part} of our altruistic motivations. It predicts that humans will be moved to greater altruistic exertions by a desire to live up to others’ trust, and hence to avoid various social penalties. This perhaps diminishes the fiduciary non-cognitivist’s biological commitments to a degree.

Virtually every M-member has the power to frustrate the altruistic desires of every other member. This is partly because altruistic-desire-frustrating actions are easy to perform. It is easy to hurt, to lie, to steal, and so on, assuming there are living creatures around that can be hurt, lied to, and so on. Also, because altruistic desires are universal, such an action performed in one part of the world will frustrate the relevant altruistic desires of moralists in every other part of the world, no matter who they are or whom they know.

Thus base conditions (i) and (ii) are met. It would potentially serve the interests of ‘proto-moralists’ to form a network of entrusted concern relationships in which altruistic desires would be weighted for both entrusted and entrusting parties. But the satisfaction of (i) and (ii) is not enough to generate a very strong prediction that they would form such a network. Being a member of such a trust network and honouring one’s obligations as such would be costly. It would mean, roughly, acting \textit{only} on the bidding of your altruistic desires, in all of the decisions in which the terms of the trust network give entrusting parties voices. Which decisions would these be? To get the

facts about our moral thought and talk right, I must postulate that the terms of the moral trust network would give entrusting parties voices in virtually all decisions that entrusted parties face, since we seem to be able to make judgements about the moral propriety of almost any act. Forming or joining the moral trust network might thus be a prohibitively demanding undertaking.

Whether it would actually serve moralists’ interests to form a network giving members such pervasive presumptive influence on one another’s decisions depends largely on the following factors: (i) the strengths of their altruistic desires relative to their non-altruistic desires; (ii) the extent to which acting in accordance with their altruistic desires would mean acting in defiance of their non-altruistic desires; (iii) the size of the group; (iv) the costs of being either a non-member or an unscrupulous member.

Regarding (iii), the size of the network matters because greater participation means, for each member, a greater satisfaction of altruistic desires in exchange for the same frustration of non-altruistic desires. Thus the larger the network would be if it were formed, the greater would be the utility of its existence. Unfortunately, however, the benefits of the network’s existence would be the same for a person whether she was a member of it or not. A non-member would enjoy the benefits of members’ efforts to refrain from frustrating altruistic desires just as a member would.

I think that I can plausibly postulate that (a), if the members of the moral trust network were to form some sort of agreement to, or were disposed to, impose heavy costs on non-members and unscrupulous members, then, as a result, it would typically be instrumentally rational for proto-moralists to join the moral trust network in order to avoid these costs; and (b), given this, it would be instrumentally rational for proto-moralists to form the moral trust network (perhaps in conjunction with an agreement to punish defectors etc.) in the first place. The costs in question might involve any sort of ill-treatment; violence, ostracism, and disparaging gossip seem likely candidates. So I shall posit this, and rely on factor (iv) to explain the rationality of forming and participating scrupulously in the moral trust network, despite the heavy sacrifices involved in being a scrupulous moralist.

126 Or rather, it gives them voices in each other’s choices of which norms to conform to in virtually all decisions that they face. But for my present point this is unimportant. Being obliged to accommodate others’ preferences in one’s choices among norms to conform to is not generally less demanding than being obliged to accommodate their preferences in one’s choices among actions.
Difficult questions still remain, of course, about why a (possibly) universe-spanning moral trust network exists. I have been artificially assuming that (proto-)moralists everywhere are able to communicate with each other, and so might have negotiated an agreement to form the moral trust network. How could it come about absent such communication? The following just-so story would explain the all-encompassing pretensions of the moral trust network, and I suspect that I am committed to the truth of something like it.

At some stage in pre-history, there existed a population of hominins from which we are all at least in part descended, who possessed altruistic desires, and who were capable of forming networks of entrusted concern relationships. The population was divided into somewhat separate but sporadically interacting and inter-breeding groups. By suitably gradual and haphazard processes, the groups formed moral trust networks, or proto-types of moral trust networks, because it was instrumentally rational for them to do so for the sorts of reasons I have described. The same sorts of considerations that favoured forming moral trust networks within a group would also have favoured forming a broader trust network spanning all of the groups. Altruistic-desire-frustrating actions performed in one group would frustrate the altruistic desires of members of other groups. If a moral trust network existed in group A and not in group B, members of A could defect to B, and so enjoy the benefits of the sacrifices made by members of A without reciprocating. So it would have helped to preserve the moral trust network in A if the members of A made life miserable for B, by making war on them etc., and so deterring defection from their own group. Thus it would have served the interests of each group to have had a moral trust network of its own, not just to inhibit frustrations of members’ altruistic desires, but also as a means of avoiding the hostility of other groups with moral trust networks.

There would, moreover, have been pressures toward homogenizing the moral trust relationships of different groups into one. First, if group B had a more lenient moral trust network than group A, requiring smaller sacrifices of its members, members of A would still have a reason to defect to B, and so A would have a reason to be hostile to B. So we should expect different moral trust networks to have tended to be similarly demanding. Second, it would have helped to minimize conflict if the discretion-licencing preferences of a moralist in B were discretion-licencing for him in A too. That way, members of A would have had fewer suspicions about whether members of B were getting away with altruistic-desire-frustrating actions that they wouldn’t have got
away with had they been in A. They could have relied on members of B to police each other’s behaviour, without feeling the need to intervene themselves. Moreover, if some act of great cruelty had been performed by a member of B, and A was poised to engage B in a crusading war to, as it were, punish them for permitting it, it would have been helpful if the same facts that convinced the members of B that the agent was acting on a discretion-licencing preference could also have been adduced to convince the members of A of the same.

Third, it would have helped to resolve conflicts if members of A had had the same discursive entitlements to demand that members of B explain the reasoning behind their preferences that members of B had, and vice versa. If a member of A were temporarily living amongst the members of B, it would have been helpful if the members of B could have scrutinized and policed his behaviour in just the same way in which the members of A could, so that if they ‘punished’ his action, they could have explained this ‘punishment’ in a way that would have satisfied members of A that this was not done out of hostility toward A. Fourth, if the members of A and B were parties to the same moral trust network, they could have more easily shared information about who was and who wasn’t unscrupulous, and so avoided misplacing their trust qua moralists.

There would presumably have been ample opportunities for members of the population to ‘experiment’ with moral trust networks spanning multiple groups. Whenever a large group split into two smaller groups, the possibility of maintaining the original trust network amongst all the members of the original group, rather than dividing it, would have presented itself. Whenever there was a lot of trade, cooperation, and back-and-forth migration between two groups, it would presumably have been convenient and easy for the members of each to extend the trust and commitments to one another that they extended to members of their own groups.127

Let us suppose that these counterfactual claims are all correct, and boldly continue our story chiefly in the indicative mood. Unsurprisingly, a moral trust network spanning the entire population eventually developed. Members of the population were able to

---

127 According to Gibbard’s summary of anthropological research on present-day hunter-gatherers, interaction among groups is common. ‘Current hunter-gatherers camp temporarily in small groups, with about twenty adults. The groups keep splitting into smaller groups and reforming in new ways, depending on such things as season and social tensions. As a result, kinship within a camp is not always close, and yet people in the camp share food and cooperate extensively. Intercamp relations are cultivated, perhaps as a kind of insurance.’ Gibbard, Wise Choices, 258n.2.
safely assume that anyone they encountered was a member of the moral trust network, whether the encountered parties were members of their own groups or strangers. They could extend their trust and commitments qua moralists indiscriminately to everyone in the world who would reciprocate, and who met a threshold of trustworthiness, and this lack of discrimination would bring with it almost no costs.

This situation persisted long enough to affect the evolutionary trajectory of the population. Natural selection favoured members who could participate scrupulously in the network with ease; who had an intuitive and unreflective facility for the sort of reasoning that produced voiceable preferences, and the specialized language with which such preferences were voiced, discussed, and scrutinized; and who were disposed to indiscriminately extend their trust and commitments qua moralists to strangers and acquaintances alike. The result was that these dispositions and this facility became at least partly innate (since having a fitness-enhancing trait innately means having it more reliably and cheaply) and ubiquitous in the population and the modern humans who descended from it. This is the sort of story I must tell to explain why the moral trust network does not seem, like many other trust networks, to be confined to any particular community, family, or organization, and why I am capable of having moral disagreements with unknown moralists on cannibal islands and on Moral Twin Earth.128 I shall elaborate a little on this story when I consider some objections to my explanation for ‘moral universality’ in Section 5.4.

128 Hare, Language of Morals, 148-9; Horgan and Timmons, ‘Troubles for New Wave Moral Semantics’, esp. 164-6. I discuss Hare’s cannibal case briefly in 5.4.4. I believe that fiduciary non-cognitivism can vindicate our intuitions about Moral Twin Earth, in spite of the fact that Twin Earthlings are beings whose evolution was separate from that of terrestrial humans, because they are described as being extremely similar to terrestrial humans. They seem to differ from us only in ‘temperament’, or in the relative strengths of some of their non-cognitive attitudes (165). Assuming that my empirical claims about entrusted concern relationships and trust networks are true with regard to humans, it is reasonable to infer from Horgan and Timmons’s description that Twin Earthlings possess the same facility for forming, and tendency to form, trust networks that humans possess. This being so, it is reasonable to infer from the description of the role played by terms like ‘wrong’ in their thought and talk that they are parties to the moral trust network. Thus fiduciary non-cognitivism vindicates an intuition of moral disagreement in this case. Fiduciary non-cognitivism might very well be unable to vindicate similar intuitions about cases involving less humanlike extraterrestrials, and may have to rely on an error theoretic explanation (attributing the intuition to an irrational human tendency to anthropomorphize, or something similar).
5.2 Non-moral Precedents for Some Novel Features of the Moral Trust Network

The moral trust network, on my theory, has several novel features. Some of these features are not apparently shared by any other trust network. Thus my postulation that the moral trust network has these features is somewhat ad hoc. However, I shall now argue that some of the novel features I wish to postulate are plausibly shared by some non-moral trust networks, if people’s capacities to have third-party disagreements in preference is an indicator of the existence of those networks. By so arguing, I hope to glean skerricks of plausibility for my postulation that the moral trust network possesses these features.

Since it is not concerned with the moral trust network specifically, this section of Chapter 5 can be regarded as a continuation of the general discussion of trust networks that appeared in Chapter 3. It makes sense to include the material below in the present chapter rather than Chapter 3, firstly because it is tangential to the central concern of that chapter, which was to argue for the existence of networks of entrusted concern relationships permitting third-party non-cognitive disagreement; and secondly because the new material is more narrowly focused on supporting the particular claims of my metaethical theory.

The novel features whose independent existence I wish to argue for are saliently exemplified by what I will call *community-specific trust networks*. These are dual-order fully connected networks whose membership robustly approximates that of a particular community, rather than, say, a family or an organization. A community-specific trust network is thus somewhat exclusive—it consists of only a proper subset of all ‘trustworthy’ people. But it may be quite large, large enough so that many of its members are strangers to each other. Its members will tend to have some sort of shared identity. The weighted desires of its members will be desires that the members of the community tend to have in common, and that outsiders are less likely to have. These will typically be desires concerning the community’s distinctive interests, including desires concerning the honouring of its traditions, the preservation of its culture and language, its engagement with outsiders, and the improvement of its reputation and welfare.

Here are some examples of communities amongst whom, I contend, we normally find community-specific trust networks: those large subsets of the populations of
democratic states who in a suitably strong sense ‘identify’ as citizens or nationals of those states; the similar but smaller subsets of the populations of non-democratic states who have some sort of influence on how they are governed and a stake in preserving the status quo; classes within states that have strong group identities; and groups unified by a common and distinctive culture or religion, with strong group identities, in states that are not culturally or religiously homogeneous. These examples should give an idea of the sense of ‘community’ that I have in mind in this discussion: it is not so much the sense that we associate with small country towns, but rather is the sense used in talk about e.g. ‘the Vietnamese community’ in a particular country.

My evidence for the existence of community-specific dual-order fully connected networks is the fact that there are disagreements in preference in which members of a certain community generally can participate, while non-members cannot. That this fact constitutes such evidence depends, of course, on the plausibility of theoretical claims I made in Chapters 2 and 3. Here’s an example of a disagreement in preferences between two members of a disadvantaged and historically oppressed black community in a democratic country with a majority white population. The speakers are not well acquainted, and do not have any shared projects, so they are presumably not parties to any personal entrusted concern relationships. (Readers may imaginatively situate the following exchange in a possible world in which the attitudes attributed to the parties do not ring too false.)

**News Anchor**

**B.** I’ve applied for a job as an anchor at the X News network.
**A.** Why do you want to work there? They’re part of the white establishment. They always present black people as criminals. You’ll be supporting them.

**B.** If people see me on TV reading the news, they’ll see a black woman who’s well-spoken, well-dressed, has a respectable job, and holds a position of authority. It might help to change people’s opinions, especially people who watch X News, whose stereotypes are rarely challenged.

**A.** Listen to yourself. ‘Well-spoken’, ‘well-dressed’! You mean by white people’s standards.

**B.** Yes, by white people’s standards. I take it you’d rather I didn’t work there.

**A.** You’re right. I’d rather a black woman didn’t throw her support behind a biased news network, in order to make black people seem more pleasing to white people.

**B.** Well I’d rather I did work there. It’s a chance to make progress in overcoming prejudice.

This seems like a disagreement. It seems like a normative disagreement, which I daresay it is. But the parties communicate their disagreeing attitudes by reporting
mutually unsatisfiable preferences, so the disagreement is a prima facie disagreement in preference, whatever else it may be. And, like all of the robust non-cognitive disagreements that I’ve discussed so far, it plausibly relies on the existence of a certain kind of relationship between the speakers. A member of the white population couldn’t chime in with her own preference report, siding with A or B, and citing the same sorts of grounds as those that A and B offer, without seeming presumptuous. Her preference would not be one that A or B would feel obliged to give consideration to, and would not seem to put her in disagreement with A and B.

There is a complication here. If a member of the white population reported a preference and gave obviously moral grounds for it, then her contribution might not seem presumptuous. Everyone, seemingly, has a voice in everyone else’s decisions, insofar as they are treated as moral decisions. So when we consult our intuitions about the presumptuousness of a voiced preference, and its capacity to put the voicer in disagreement with others, for evidence about the existence of non-moral trust networks, we must be careful about how we characterize the grounds that the voicer gives for her preference.

Our intuitions about News Anchor support the hypothesis that there is an entrusted concern relationship between the interlocutors. They don’t specifically support the claim that the interlocutors are members of a dual-order fully connected network. But we can get more specific evidence by imaginatively elaborating on the case. For instance, if we imagine other members of the relevant black community chiming in with their own preferences, and if the preferences have the right sorts of ‘community-specific’ grounds, their contributions will not seem presumptuous, and they will seem to disagree with the parties whose preferences oppose theirs. This is reason to think that A and B are part of a network of trust relationships with a common set of weighted desires rather than a mere dyadic entrusted concern relationship. And since some of the disagreements will be third-party disagreements, this, together with the arguments I gave for the Four Relationships Hypothesis in 2.2.2, give us reason to think that the interlocutors are parties to second-order entrusted concern relationships as well as first-order ones.

\[129\] Cf. Philippa Foot’s observation that, in contrast to other kinds of approval, ‘anyone can approve or disapprove’ on grounds that ‘come within the sphere of public manners and morals’ (Virtues and Vices, 192). The present section borrows from Foot’s account of the conditions for approval and disapproval. I discuss this account in Section 3.1.
Finally, we can imagine the discussion straying from the practical decision at hand to cases in which any other members of the black community face decisions similar to those of B. (It is tempting to imagine discussions of counterfactual decisions here, though perhaps we had better not, as I have not given an account of non-cognitive disagreements about counterfactual decisions.) It is intuitively plausible that parties who disagreed about the original case would disagree about the new cases too, if those cases were sufficiently similar to the original case. This suggests that the network of trust relationships of which A and B are members is not merely one in which B is the sole common first-order entrusted party, but rather is one in which every member is an entrusted party in an entrusted concern relationship with every other member. If we have all of these intuitions, then the hypothesis that A and B are members of a community-specific dual-order fully connected network is probably the simplest way to account for them.

The property of community-specific trust networks that explains their possession of the novel features that I am interested in is that their members need not all have met or communicated with one another. Community-specific trust networks are typically, in part, networks of strangers. Thus their members will typically have entrusted concerns to, and be committed to accommodate the preferences of, complete strangers.

One consequence of this is that we should expect these trust networks to have what I shall call ‘standardized default terms’. Since the weighted desire-types for all parties in a dual-order fully connected network are the same, and since in each of the first- and second-order networks the admissible ground-types are the same among all members, it is easy to imagine that the terms of the relationships in some dual-order fully connected networks might be ‘standardized’. That is, the terms of the relationships in each of the first- and second-order networks might all be identical. Thus, for instance, there would be no difference in the extent to which one party strained another’s trust by resisting her preferences (and if one entrusted party in the first-order network could make what in Appendix B I call ‘conflict excuses’, then all could). This is more plausible in the case of community-specific trust networks, because any two members who have never met each other will not have had the chance to negotiate special terms for their shared relationships in the network. And it seems unavoidable that there should be at least some standardized default terms, which every member of the network must (tacitly) know, and which must determine the sort of trust and commitments that she extends to strangers in the network. (It is of course compatible with this hypothesis that
the network also includes entrusted concern relationships whose terms deviate somewhat from the default.) I shall posit that the moral trust network shares with community-specific trust networks the property of having standardized default terms.

If members of a trust network don’t know each other, the trust and commitments they extend to each other must be generic—by which I mean undifferentiated, save perhaps in degree, from the trust and commitments that they extend to other unknown members—and impersonal. The trust relationships that they form by reciprocating each other’s generic and impersonal trust and commitments must also be generic and impersonal. There is nothing surprising about the idea of impersonal trust relationships whose terms are determined by convention rather than negotiation. We trust the strangers around us to obey conventional rules all the time. When I drive I trust my fellow motorists to obey traffic laws. (This sort of trust is generic as well as impersonal—I trust all unknown motorists alike to follow the same rules.) That actual entrusted concern relationships are ever impersonal is a more surprising supposition. This is because an entrusted concern relationship is a fairly complicated sort of trust relationship; and because it obliges one party to be responsive to the mental states of the other, unless his own mental states licence him to do otherwise, and this sensitivity of trust-based obligations to publicly inaccessible mental states makes conformity to those obligations difficult to police. But our reasons for believing that community-specific trust networks exist are reasons for believing that impersonal entrusted concern relationships exist. I shall posit that the entrusted concern relationships among members of the moral trust network as such are (predominantly if not uniformly) generic and impersonal. Up to a point, the more plausible it is that community-specific trust networks exist, the more plausible this postulation is.

The possibility of fully connected networks of entrusted concern relationships some of whose members are strangers to each other raises a question about what could qualify a person for membership in the first place, if this qualification potentially earns one the trust of perfect strangers. There seem to be two main possibilities, as well as arrangements involving elements of both. First, it might suffice for membership in the whole network to be accepted into it by a proper subset of its members, at the discretion

---

130 By ‘degree’ I mean degree of trust and strength of commitment. I am inclined to say that generic trust is trust in each member of a group of people to do the same thing or adhere to the same rules, whether or not the degree of trust is uniform across the group; and that generic commitment is commitment to each member of a group of people to do the same thing, whether or not the strength of commitment is uniform across the group.
of that subset, assuming the would-be member reciprocates the trust of all the other members. Second, there may be some ‘objective’ criteria, specified by the terms of the trust network, such that if a person meets them, and is sufficiently trustworthy, and is willing to reciprocate the trust and commitments to accommodation of all other members, she automatically becomes a member of the network and enjoys the trust of all its members.

I am going to postulate that membership of the moral trust network is determined wholly in the latter way. It will improve the plausibility of this postulation if I can give a real-world non-moral precedent for this sort of arrangement. If they exist, community-specific trust networks provide such a precedent, for people become members of these networks at least partly by meeting certain criteria, rather than merely by being admitted into the network by some of its members. Members of a country’s ‘Greek community’, and hence of the corresponding trust network, must be of Greek descent, and have some cultural ties with Greece. If an Australian voices a preference about whether Australia should become a republic (or rather, about whether certain people should take steps to bring this about), on grounds of national pride or tradition or the national interest, no one can have a voiceable preference, based on non-moral grounds, that puts him in disagreement with her unless he too is an Australian citizen.

The moral trust network would be still less exotic if I could show that there were community-specific networks whose memberships depended only on their satisfying certain ‘objective’ criteria (and making the requisite commitments etc.). Unfortunately, however, it does not seem implausible to me that some sort of voluntary acceptance or acknowledgement by members of a community-specific trust network is always a necessary condition for membership of such a network. If this is right, then my postulation that the membership of the moral trust network does not depend on any such acceptance is, I admit, regrettably novel.

---

131 The trustworthiness and ‘willingness to reciprocate trust and commitment’ criteria may or may not be among the criteria specified by the terms of the trust network, depending on whether or not it is within the scope of the terms of a fully connected network of entrusted concern relationships to specify such things.
5.3 Independent Evidence for the Existence of the Moral Trust Network

5.3.1 Philippa Foot on Approval and Disapproval

We have a reason to posit the moral trust network that is independent of the plausibility of non-cognitivism, and the slender evidence that I have adduced for its existence so far in the course of arguing for the existence of trust networks with similar properties. As I shall argue, we must posit something like the moral trust network if we are to explain the phenomena of moral approval and disapproval. In what follows I will sometimes use the term approbation as an umbrella term for approval and disapproval, and approbative attitude as an umbrella term for an attitude of approval or disapproval.

Philippa Foot argues that approval and disapproval can, as a general matter, only exist in an appropriate social context. This is at least very plausible in non-moral cases. Parents may disapprove of their daughter’s marriage on non-moral grounds, for instance, ‘because the man is too old for her, or not rich enough, or not well enough connected; they may disapprove because they think that the marriage will not work out well, or because the family’s honour or pride is at stake.’ A stranger may have the same ‘opinions’, yet cannot disapprove of the marriage on non-moral grounds, even if she genuinely cares ‘about the girl’s fortunes, or the fortunes of the family.’ If ‘an elderly couple are thinking of retiring to some seaside resort, and the evidence is against the success of the venture’, a ‘relative or friend … may therefore disapprove of the idea’, but a stranger or a neighbour may not, even if she cares more about the couple’s welfare than their friends and relatives do. Participants ‘in a particular enterprise, like giving a dinner party, or robbing a bank’, may have a capacity to approve or disapprove of actions aimed at furthering the goals of the enterprise, which outsiders lack. There are policies of countries that people can only approve or disapprove of if they are citizens of those countries, unless they do so on moral grounds, or possess relevant recognized expertise. Foot argues that the attitudes of favouring, being for, and being against, are similarly dependent on social context. A stranger, she observes, would be

Foot, Virtues and Vices, 192-3.
Foot, Virtues and Vices, 193.
Foot, Virtues and Vices, 194.
Foot, Virtues and Vices, 197.
Foot, Virtues and Vices, 200-202.
presumptuous if he said to the elderly couple who were thinking of moving to the seaside, ‘I’m against your going’.\textsuperscript{137}

Some of Foot’s examples, and her general thesis, may remind readers of my examples and thesis in Chapter 1 regarding robust non-cognitive disagreement. This is no accident. My explanation for the robust non-cognitive disagreements described by Stevenson, and my use of examples involving ‘well-wishers’, were directly inspired by Foot’s examples and argument. My account of the conditions for non-cognitive disagreement owes a debt to Foot’s account of the social conditions for approval and disapproval. She holds ‘that approval and disapproval can, logically speaking, exist only against a background of agreement about the part that other people’s views shall be given in decision-making.’ To approve or disapprove of some option in a decision, it must be recognized that one is entitled to be listened to\textsuperscript{138}—that one ‘has a voice in the matter’.\textsuperscript{139} My claim that preferences robustly disagree only if they are based on the right kinds of desires (evidenced by the grounds given for them) is another echo. According to Foot, even when a social setting exists which permits a person to approve or disapprove of an option on non-moral grounds, she may still only be able to approve or disapprove on grounds of a specific sort. She writes:

A parent cannot disapprove of the marriage of one of his children on the grounds that a neighbour, with no particular standing in the matter, will be annoyed by it .... If he disapproves he must hold an opinion about the way the marriage will work out for his son or daughter, or about the way the family fortunes will be affected by it.\textsuperscript{140}

My Entrusted Concerns Hypothesis, and more particularly the Reason-to-Consider Hypothesis of which the former hypothesis is a precisification, draw on Foot’s diagnosis of approval and disapproval. My account of the conditions for non-cognitive disagreement is more specific than Foot’s account of the conditions for approval and disapproval, however. Foot says nothing about trust.

If, as I am persuaded, non-moral approbation requires some kind of social relationship between the approver and the approved, then we must assume that moral approbation does too, or explain why it does not. What would this require? According to Foot, there must be ‘social practices’ establishing ‘that anyone is to listen to anyone

\textsuperscript{137} Foot, \textit{Virtues and Vices}, 200.
\textsuperscript{138} Foot, \textit{Virtues and Vices}, 198-9.
\textsuperscript{139} Foot, \textit{Virtues and Vices}, 203 (Foot is discussing moral approval here; cf. 201 for a similar usage in connection with favouring and being for).
\textsuperscript{140} Foot, \textit{Virtues and Vices}, 194.
when considerations are brought forward which are moral considerations.\textsuperscript{141} It is perhaps ‘the custom of every society in the world, to take matters such as killing, stealing, and lying as a concern of the community. Everyone has a voice in the matter of whether or not these things are to be done.’\textsuperscript{142} Because the existence of specific ‘social practices’ and ‘customs’ is a rather contingent matter, and because they tend to distribute entitlements parochially within a particular society or culture, Foot has difficulty explaining how it is that we can morally disapprove of actions performed by members of communities that have had little or no contact with our own, and why we are confident that they can morally disapprove of our actions.

5.3.2 \textit{Reason to Think that Entrusted Concern Relationships Enable Approbation}

It is a virtue of fiduciary non-cognitivism’s otherwise embarrassing postulation of a universe-spanning moral trust network that it can explain these phenomena, as I will now argue. Entrusted concern relationships, as I shall argue in a moment, provide the ‘social context’ that approval and disapproval require. If the moral trust network exists, then every moralist shares entrusted concern relationships (at least four) with every other moralist. So fiduciary non-cognitivism explains the existence of the social contexts that enable moral approval and disapproval. Moreover, despite the implausibility of positing the moral trust network, the postulation is not utterly implausible, and it is hard to think of another way of explaining the phenomena that is very dissimilar from that offered by fiduciary non-cognitivism that is any more plausible. So the postulation of the moral trust network is, all things considered, not nearly as implausible as it at first appears. (Note, however, that the ‘second-order’ of the moral trust network is not needed to explain moral approval; fiduciary non-cognitivism therefore incurs a special cost by positing it.)

I am not going to defend a complete analysis of approval and disapproval. But I do need to say something about the nature of these states if I am to claim that entrusted concern relationships plausibly provide the social contexts that make approval and disapproval possible. To start with, I propose that approval and disapproval are attitudes of entrusting parties, not entrusted parties. An entrusted party as such can’t approve or disapprove, because he only has a voice in his own decisions, and approval and

\textsuperscript{141} Foot, \textit{Virtues and Vices}, 202.
\textsuperscript{142} Foot, \textit{Virtues and Vices}, 203.
disapproval seem to be directed at other people and their actions. I further propose the following first-pass analyses:

**Approval (first pass).** An entrusting party approves of an option φ if and only if she has a voiceable preference that her partner should φ, and does not intend to yield or waive in the matter of his φ-ing.

**Disapproval (first-pass).** An entrusting party disapproves of an option φ if and only if she has a voiceable preference that her partner should not-φ, and does not intend to yield or waive in the matter of his φ-ing.

These first-pass analyses contradict a datum that Baker and Woods have pointed to concerning consistent combinations of approval and disapproval states. Those writers consider the following examples:

(1a) Joe disapproves of dancing.
(1b) Joe disapproves of not dancing.
(2a) Joe disapproves of dancing.
(2b) Joe approves of dancing.

According to Baker and Woods, ‘It should be clear that the pair (1a) and (1b) is not discordant [i.e., roughly, inconsistent]; it is simply the attitudes of someone who is mean spirited. The pair (2a) and (2b) is, however, discordant.’ However, if my partial hypothesis is correct, we should not intuit this difference between (1a)-(1b) and (2b)-(2b). These sentences would merely imply the following about Joe’s mental state:

(1a’) Joe has a voiceable preference for not dancing.
(1b’) Joe has a voiceable preference for dancing.
(2a’) Joe has a voiceable preference for not dancing.
(2b’) Joe has a voiceable preference for dancing.

While I share the intuition that Baker and Woods report, I am not sure what to make of it. Suppose that a curmudgeonly person at a dance expressed disapproval of dancing and disapproval of not dancing, and someone asked ‘What would you have us do?’ I can think of nothing sensible she might say that wouldn’t involve contradicting or reversing one of her expressions of disapproval. Or rather, not quite anything: I can imagine her saying, ‘Well since you’ve gone to the trouble of organizing this frivolous event, you’d better dance. But you should never have organized it in the first place.’ If she said this, we would be allowed to interpret her as disapproving of dancing in a sense, and disapproving of not dancing in a (perhaps different) sense. And it would

---

seemingly be odd to suppose that she approves of dancing in any sense. So, apparently, disapproving is more easily done than approving. Approval of an action $\phi$ in a choice situation $C$ requires approval (or at least non-disapproval) of the choices that put the agent in $C$; whereas it suffices for disapproval of $\phi$ that one disapproves of the choices that put the agent in $C$. At least, this seems to be true of some cases.

It would be beside my purpose to try to give a thorough account of this backward-looking character of approbation. I shall instead confine my attention to cases in which ‘historical’ factors do not intrude, e.g. cases in which the agent is not in her choice situation by choice. As far as I can see, my first-pass ‘accommodation-obliging preference’ analysis does not fall afoul of intuitions of the sort that Baker and Woods elicit in these cases.

However, it is defective for another reason. Observe that it does not sound quite right to say that the parties in Stevenson’s restaurant case approve of their preferred options, and it sounds wrong to say that they disapprove of their dispreferred options. Yet my first-pass analysis implies that they have these approbative attitudes. I mentioned above that Foot thinks that attitudes of favouring as well as approval depend on ‘social practices’; this, I suggest, is a case in which the parties’ entrusted concern relationship permits them to favour options. But their accommodation-obliging preferences do not amount to attitudes of approval and disapproval.

As I observed in 1.1.4, it is more plausible to say that the parties in Stevenson’s restaurant case are making requests of each other than that they are prescribing their preferred options. This may partly explain why they do not amount to approbative attitudes. Approval and disapproval have a prescriptive flavour. They are not attitudes we associate with mere requesting. However, it would be unsatisfactorily vague and superficial to analyse approving as ‘voiceably preferring with a prescriptive flavour’. Let us see if we can do better.

What explains why the preferences voiced in Stevenson’s restaurant case seem to be mere favouring, and not approbative attitudes? Here are three hypotheses:

(1) Voiceably preferring is approving only in ‘non-joint-action contexts’, i.e., only when the entrusting party’s voiceable preference is not for an option in a decision about how she and her partner are to coordinate their actions in pursuit of a common goal. Since the context in which the preferences voiced in Restaurant is a ‘joint-action context’, this hypothesis would explain why neither party disapproves of the other’s preferred option.
(2) Voiceably preferring is approving only when the entrusted party has a sufficiently strong trust-based obligation to accommodate the entrusting party’s preferences. The parties in Restaurant presumably have only weak trust-based obligations to accommodate each other’s preferences, so their accommodation-obliging preferences don’t, as we might intuitively say, have the force and gravity of approbative attitudes.

(3) Voiceably preferring is approving only when the entrusting party’s preference doesn’t depend on weighted desires that she and her partner do not share (i.e., desires that would not be satisfied by the same outcomes). Since each party in Restaurant is voicing a preference based on desires that she does not share with her partner, neither preference amounts to an approbative attitude.

This roughly corresponds to Foot’s own proposed explanation for the difference between approving and favouring: cases of merely favouring or merely ‘being for’, rather than approving, are ones in which the relevant social arrangements that entitle the relevant parties to voices do not have the exclusive purpose of furthering some common ends, and in fact give each potential favourer a voice in the other parties’ decisions on private and self-regarding grounds. Foot, however, also thinks that approval and favouring involve different kinds of internal mental states: approval involves having an ‘opinion’ that is to be given consideration, favouring involves having a ‘wish’ that is to be given consideration.

Among hypotheses (1)-(3), I think that (3) is the most plausible. Against (2), observe that I, an Australian, can disapprove of the actions of an Australian prime minister on non-moral grounds (perhaps with respect to a ‘national identity’ issue, e.g., whether Australia should become a republic). Yet the prime minister’s trust-based obligation, qua member of the community-specific trust network consisting of the Australian citizenry, to accommodate my preferences is surely very weak indeed.

Here is a case that tells against (1) but supports (3). Suppose that A and B have agreed to go to a movie after dinner, and A voices a preference that they should dine at En Route, a restaurant which they will pass on their way to the cinema, because she doesn’t want them to be late for the movie. B shares this concern. But he nonetheless prefers that they dine at Détour, because it serves one of his (B’s) favourite dishes, even

---

144 Foot, Virtues and Vices, 200-201.
145 Foot, Virtues and Vices, 201-202.
though getting to that restaurant would take them out of their way, and so delay them. It seems fairly plausible that A could disapprove of going to Détour, given the grounds for her preference, and very plausible that B could not disapprove of going to En Route, given the self-regarding grounds of his preference. A favours going to En Route; B favours going to Détour; but A and only A might disapprove of her dispreferred option. This case undermines (1) because it suggests that approbation is possible in a joint-action context. It supports (3) because it is a case in which the fact that an option serves ends that the parties are presumed to share seems to explain why a preference for that option can put its possessor in a state of disapproval.

If (3) is correct, the version of fiduciary non-cognitivism that I defend seems to be well positioned to explain why moralists as such are capable of approving and disapproving of actions rather than merely favouring or disfavouring them. For the voiceable preferences of moralists are based chiefly on altruistic desires, which all may be presumed to share in varying degrees, so they are preferences for options that tend to serve shared ends.

Setting aside the ‘backward-looking’ conditions for approval and disapproval, I propose that the following is approximately true:

Approval (second-pass). A person A approves of a person B’s φ-ing in circumstances C just in case A has an accommodation-obliging preference that B should φ in C, and A has not yielded or waived in the matter of B’s φ-ing in C, and A’s preference is explained by the fact that there is some subset D of A’s weighted desires such that

(i) D suffices to explain A’s preference that B should φ in C, and
(ii) the desires in D fall under desire-types that are weighted for B as well as A, and
(iii) the desires in D are ones that A and B have in common (albeit perhaps differing in strength).

Disapproval (second-pass). A person A disapproves of a person B’s φ-ing in circumstances C just in case A has an accommodation-obliging preference that B should not-φ in C, and A has not yielded or waived in the matter of B’s φ-ing in C, and A’s preference is explained by the fact that there is some subset D of A’s weighted desires such that

(i) D suffices to explain A’s preference that B should not-φ in C, and
(ii) the desires in D fall under desire-types that are weighted for B as well as A, and
(iii) the desires in D are ones that A and B have in common (albeit perhaps differing in strength).

Condition (iii) in each of these partial analyses should perhaps say that the desires in D are ones that A and B recognize that they have in common, or more strongly, that there is a mutual understanding between A and B that satisfying the desires in D would serve ends that they share, or some other variation on this idea.
This account of approval and disapproval is incomplete, not just because it ignores ‘backward-looking’ conditions, or because of my indecision about how condition (iii) should be formulated, but also because I have devised it without taking into account the complications that may arise when we consider the relationships between morally approving or disapproving of φ-ing on the one hand (a two-way distinction) and, on the other, judging that φ-ing is morally required, permissible, or impermissible (a three-way distinction). I am not sure whether the account makes the right predictions about these relationships, in part because I am not sure what I my account ought to predict. Plausibly, wrongness judgements ought to be correlated with disapproval and requirement judgements ought normally to imply approval. But what about mere permissibility judgements? I shall set these issues aside, however, and content myself with having provided the beginnings of a satisfactory account of approval and disapproval.

The account, incomplete though it is, explains, or at least is not challenged by, most of our intuitions about the cases I have discussed so far. For a start, its prediction that an approving or disapproving party must be the entrusting party in a relevant entrusted concern relationship with the disapproved party seems accurate. The parties in Flat (Preferences), Retail, Rival Hostesses, Football (John and Well-Wisher), and Restaurant (Blackmail I) presumably do not share relevant entrusted concern relationships, and do not seem to be voicing approbative attitudes (or attitudes of favouring for that matter).\(^\text{146}\)

The following are cases in which an entrusted concern relationship between the parties that would make their reported preferences voiceable plausibly exists: Restaurant, Invitation, Football (Mother and Son), Drinks, Marriage (Imperatives), Football (Parents I), Football (Parents II), Photographs, and News Anchor.\(^\text{147}\) Among these, those in which one or more parties might be thought to be voicing approval or disapproval are Football (Mother and Son), Football (Parents I), and News Anchor, and possibly also Invitation and Marriage (Imperatives). In Football (Mother and Son) and Football (Parents I), John’s mother plausibly disapproves of John’s playing football because of its dangers, and in the latter John’s father plausibly disapproves of John’s

\(^{146}\) Flat (Preferences) and Retail can be found in 1.1.2, Rival Hostesses in 1.1.3, Football (John and Well-Wisher) in the introduction to 1.2, and Restaurant (Blackmail I) in 1.2.4.

\(^{147}\) Restaurant and Invitation can be found in 1.1.1, Football (Mother and Son) in 1.1.6, Drinks and Marriage (Imperatives) in 1.2.5, Football (Parents I) in the introduction to Chapter 2, Football (Parents II) in 2.1, Photographs in 3.2.1, and News Anchor in 5.2.
quitting the football team because doing so defies a family tradition. In *News Anchor*, A plausibly disapproves of B’s becoming a newsreader because she takes X News to be part of the white establishment. And the preferences these entrusting parties report are based on desires that they expect their relevant partners to share.

It is only somewhat plausible that the parties in *Invitation* approve of their preferred options and disapprove of their dispreferred options. It is more plausible that Mr. A disapproves of inviting Toff than that Mrs. A disapproves of inviting Chum, perhaps because we detect a tincture of moral judgement in Mr. A’s reported preference for staying loyal to an old friend. I suggest that we are not much inclined to think that Mrs. A disapproves of inviting Mr. Chum because we suspect that Mr. A does not share Mrs. A’s social aspirations, and thus we are not inclined to suppose that the pair take rising in society to be a common end. Mrs. A’s preference thus has more of a favouring than an approbative character. We have a stronger inclination to think that both Mother and Aunt in *Marriage (Imperatives)* have approbative attitudes, because we take their entrusted concern relationships with Penny to be part of a familial trust network in which (it is generally assumed that) members have the common ends of both improving the family’s social standing and preserving its traditions. We would, it seems, have a stronger intuition that the parties in *Invitation* have approbative attitudes if it were suggested that their trust relationship was part of a familial network. Since we are allowed to think of it as a merely personal trust relationship, we are allowed to suppose that the parties will have weighted desires of a more idiosyncratic and self-regarding nature.

The parties in *Restaurant, Drinks,* and *Football (Parents II)* seem obviously to be voicing preferences based on private (though weighted) desires. They thus seem to merely favour their preferred options. In *Football (Parents II)*, John’s parents seem to be voicing preferences qua parties to a private entrusted concern relationship that is not part of any trust network of which John is a member. Although the parties voice preferences ostensibly concerning what John is to do, we must suppose that they are in fact elliptically favouring options that they themselves can choose, presumably options of encouraging John to choose one way or another.

*Photographs* is a difficult case. It seems to me that both Cousin Agnes and Cousin Augusta both favour and approve of their preferred options, and that Cousin Agnes disapproves of Bertie’s sending the photographs to the press, because she thinks that his doing so would tarnish the Seymour name. But it does not seem to me that Aunt
Augusta disapproves of the option *Bertie does not send the photos to the press*, even though this option would prevent the destruction of a family tradition. Curiously, I think that this is because it does not seem that Bertie could have a sufficiently strong kind of trust-based obligation to send the photographs to the press when this would ruin the reputation of a fellow Seymour. Augusta’s preference seems to have the flavour of a request rather than a demand, or the flavour of a judgement of supererogatory value rather than a judgement of perfect duty.

If my intuition on this point is not idiosyncratic, it reveals that my characterization of the network of trust relationships among the Seymours was too simple, and that I have underestimated the complex normative character of the demands and pressures that the Seymours can make of and impose on each other, even when we ignore their intimate and personal relationships. Complicating my account of the trust network among the Seymours so that it will fully accommodate my intuitions about disapproval is more than I can attempt in this work. Does the intuition about Augusta’s preference reveal a defect in my analysis of approbation? None, I think, that I have not already acknowledged. My analysis only gets the wrong result here because, as a general matter, it is not equal to the challenge of explaining the correlations between approbative attitudes and the more complex kinds of normative judgement (involving permission, and in this case perhaps, supererogation).

5.3.3 *That Something Like the Moral Trust Network is Needed to Explain Moral Approbation*

Although the analysis of approval and disapproval that I have offered is only half baked, its match with our intuitions is auspicious. This intuitive support licences optimism that a true analysis of approval and disapproval, according to which entrusted concern relationships constitute the social-contextual ingredients for these attitudes, is waiting to be discovered. Moreover, it seems that such an analysis would provide the basis for an explanation of the phenomena of *moral* approval and disapproval that was at least as plausible as the explanation allowed by any rival analysis, if that rival analysis recognized social contexts as necessary for approval and disapproval. For the ‘social context’ that would be necessary for moral approbation must be one that entitles every moralist in the universe to a voice in every other moralist’s decisions. And among those arrangements of social institutions that would provide this social context, the first order
of fiduciary non-cognitivism’s moral trust network is one whose existence can be posited with a *comparatively* high plausibility.

Here are three propositions and an argument.

(1) I can apparently morally disapprove of the actions of a moralist on Approbia, a planet in a distant galaxy that has never been in contact with Earth.

(2) A moralist on Approbia can apparently disapprove of my actions.

(3) Person A can only approve or disapprove of person B’s φ-ing if social practices or customs or trust relationships exist entitling A to a voice in B’s decision to φ or not-φ.

((3) is a disjunction of Foot’s and my proposed necessary conditions for approval and disapproval.) Fiduciary non-cognitivists can account for (1)-(3) because they posit a necessary connection between being a moralist—even an Approbian moralist—and being a party to a network of trust relationships that gives members a voice in the decisions of all other moralists. Cognitivists (and traditional non-cognitivists) can’t account for (1)-(3) unless they posit something similar.

I don’t claim that cognitivists need to make the same postulation (i.e. a universe-spanning fully connected network, and a necessary connection between being a maker of moral judgements and being a member of such a network) in order to explain (1) and (2) given (3). They could, for instance, postulate that there is a universe-spanning social contract, entitling every moralist to a voice in every other moralist’s relevant decisions (or that our approval attributions assume this). Or they could postulate that there are *local* conventions in every society in the universe conferring similar entitlements (or that our approval-attributions assume this). All I claim is that they have to postulate something like this in order to give a completely satisfactory explanation of (1) and (2), and that none of the cognitivist-friendly alternatives to my proposed network of trust relationships is significantly less costly.

Postulating the network of trust relationships required for the non-cognitivist to give a wholly satisfactory explanation of moral disagreement seemingly does not require postulating anything more extravagant than what everyone in metaethics must postulate to give a wholly satisfactory explanation of moral approbation. Does this mean that the there is no net cost for fiduciary non-cognitivism in postulating the first order of the moral trust network? Probably not. Cognitivists probably won’t want to give a completely satisfactory explanation of moral approbation if it means making such a postulation. They might well do better to simply reject Foot’s intuitive claims, or postulate that ‘approval’ and ‘disapproval’ are used in different senses in moral and
non-moral contexts, or insist that some sort of counterfactual social contract suffices to make approbation possible. But the plausibility of (1)-(3), and the obligation of all metaethicists to explain (1) and (2) given (3) if they are true, greatly lightens the burden of positing the moral trust network.

Michael Smith thinks that we can satisfactorily explain moral approbation while rejecting (3). Smith presents an account of approval and disapproval that, like mine, was directly inspired by Foot’s account and the striking observations that she adduces in support of it. Like Foot, he sets out to explain what it is that distinguishes approving from attitudes like wanting and liking. Here is his proposal:

to say that I disapprove of your behaviour, as opposed to merely dislike it, signals the fact that, as I see it, your behaviour transgresses the standards in terms of which you and I both acknowledge your behaviour is to be judged. In other words, disapproval presupposes that your behaviour is contrary to my legitimate expectations; my belief about how you will behave. Disliking your behaviour presupposes no such thing.¹⁴⁸

One party must have ‘legitimate expectations’ about how another will act if she is to disapprove of his actions. What exactly is a ‘legitimate expectation’? This much is clear: ‘expectation’ is not meant in a normative sense. Smith explicitly rules out a normative interpretation.¹⁴⁹ A legitimate expectation regarding your behaviour is, as Smith says in the quote above, just a (suitably grounded) ‘belief about how you will behave’. (Elsewhere he implies that the prediction is about how you will behave ‘at least all things being equal’.¹⁵⁰) Where do ‘legitimate expectations’ come from? On Smith’s view, social conventions can give rise to ‘legitimate expectations’ about how others will behave, but so can the requirements of rationality, ‘[f]or we all expect of each other that we will decide what to do on rational grounds’.¹⁵¹ In the non-moral cases that Foot drew our attention to, it is social conventions which explain the possibility of disapproval. In moral cases, the requirements of rationality (the ‘expansive’ sort of rationality favoured by moral rationalists) explain the possibility of disapproval. They do so because the moral rationalist’s thesis that moral requirements just are requirements of rationality is correct.¹⁵² Thus we can spare ourselves the embarrassment

¹⁴⁸ Smith actually presents this as an interpretation of Foot, though he evidently endorses it himself. The Moral Problem, 89.
¹⁵⁰ Smith, The Moral Problem, 87.
¹⁵¹ Smith, The Moral Problem, 89.
¹⁵² Smith, The Moral Problem, 85, 90.
of postulating that moralists are all parties to an agreement of some sort by embracing moral rationalism.\footnote{153}{Smith, The Moral Problem, 90.}

I am not sure how to interpret Smith’s claim that ‘my’ disapproval of ‘your’ behaviour requires that ‘as I see it, your behaviour transgresses the standards in terms of which you and I both acknowledge your behaviour is to be judged’. I am not sure, for instance, what Smith means by ‘acknowledge’, or whether ‘is to be judged’ means ‘ought to be judged’ or ‘is planned to be judged’, or whom the judge is supposed to be. But the idea of predicted conformity to ‘commonly acknowledged standards’\footnote{154}{This phrase appears in Smith, The Moral Problem, 90.} must be important to Smith’s explanation. He would fail to explain Foot’s observations if he merely proposed the following:

P. What distinguishes disapproval of S’s φ-ing from simple dislike of S’s φ-ing is merely that disapproval involves (justifiedly) predicting that S will, all things being equal, conform to standards forbidding φ.

Suppose that a young woman is planning to marry, and that one of her parents disapproves on the ground that the man’s poverty would embarrass their family. Intuitively, a benevolent stranger with the family’s best interests at heart cannot disapprove of the woman’s choice on the same ground. Presumably, if the truth of P explains our intuitions that the parent can and the stranger cannot so disapprove, it must be the case that we assume that the parent can, and the stranger cannot, justifiedly predict that the daughter would, all things being equal, follow standards forbidding marrying a man whose poverty would embarrass the family. But there is no reason I can think of why the stranger could not justifiedly make such a prediction, or why we would assume that she couldn’t. So presumably Smith rejects P.

Everything hinges, then, on the nature of the common acknowledgement condition. But Smith’s characterization of it is underspecified in multiple respects. I cannot easily see a way to precisify it that wouldn’t either fail to explain Foot’s observations or create pressure to posit that moral approbation requires a social context. And I don’t think I am obliged to evaluate Smith’s account of approval for every remotely plausible precisification of the common acknowledgement condition.

I can perhaps give an idea of the general difficulty facing attempts to provide analyses of approbation that neither contradict Foot’s observations nor imply that moral
approbation requires a social context, by evaluating the following much simpler proposal:

Q. What distinguishes disapproval of S’s φ-ing from simple dislike of S’s φ-ing is merely that disapproval involves judging that S ought not to φ.

Q is underspecified, but manageably so. It merely omits to specify the kind of ought judgement that is involved in disapproval.

If Q is to explain why the parent can and the stranger cannot disapprove in the marriage case I described above, it must be the case that the parent can and the stranger cannot judge that the young woman ought not to marry the man (presumably on the grounds that his poverty would embarrass the family), in whatever sense of ‘ought’ is relevant. There is, as far as I can see, no way to explain this difference that does not make the parent’s capacity to make the ought judgement dependent on her relationship with the young woman. So there is an approbation-enabling kind of ought judgement that is relationship-dependent. Now we must ask: if we know that one kind of approbation-enabling ought judgement requires a certain relationship between the judge and the agent whose actions are the object of the judgement, what grounds do we have for thinking that all kinds of approbation-enabling ought judgements do not require such a relationship? The fact that moral ought judgements, for instance, are approbation-enabling and that it would be far-fetched to posit social relationships of the relevant sort linking all pairs of moral judges and morally judged parties is certainly a good ground. But if this is the only ground that can be offered, the explanatory advantage that I have claimed for fiduciary non-cognitivism concerning approbation is secured. There is apparently no way of cashing out Q that allows us to provide a similarly parsimonious and unified explanation of approval (or the judgements that enable it) which does not also posit something like fiduciary non-cognitivism’s moral trust network.

5.4 Objections

I’ll now consider some objections to my explanation for moral universality. This explanation involved the postulation of entrusted concern relationships linking all agents who met certain criteria—criteria that were supposed to be satisfied by all and only those whom we take to be capable of making moral judgements, among other things. The first three are versions of the objection that the postulation of some of the required trust relationships is deeply implausible. They contend that there are possible moral disagreements (or ‘approver-approvee’ pairs) whose parties do not plausibly trust
each other in the required way. The first objection concerns disagreements one of whose parties falsely believes that the other does not meet the ‘objective criteria’ required for membership in the network. The second concerns disagreements one of whose parties has demonstrated extreme untrustworthiness through extreme unscrupulousness (e.g. inveterate hypocrisy). The third concerns disagreements whose parties share extremely antagonistic relationships (e.g. a Nazi commandant and a Jew in Auschwitz). The postulation of the moral trust network is costly even if these objections are wholly unsuccessful, so the objections must be understood as charging the postulation with a greater degree of implausibility than I have acknowledged. The fourth objection concerns the possibility of moral disagreement involving agents who lack altruistic desires.

5.4.1 Lack of Trust Owing to False Beliefs

The first objection that I’ll consider is by no means the most pressing. But it is a simple challenge that I can address fairly quickly, and doing so gives me an occasion to make my position a bit more precise, and eliminate possible sources of confusion.

Members of the moral trust network posited by fiduciary non-cognitivism extend their trust indiscriminately to everyone who meets certain criteria specified by the terms of the moral trust network. Not everything meets these criteria: non-human animals, Daleks,155 psychopaths, and unsocialized feral humans, and others who do not ‘get’ morality,156 probably do not. This seems to mean, according to fiduciary non-cognitivism, that such creatures cannot make moral judgements or have moral approbative attitudes, that they cannot be parties to moral disagreements, that they cannot be (morally) disapproved of, and that their token actions cannot be judged morally wrong or disapproved of.157 It is not obvious to me that this implication creates any problems for fiduciary non-cognitivism. But suppose that moralist A judges that abortion was wrong, that moralist B judges that abortion was permissible, and that A falsely believes that B did not meet the criteria. (Perhaps B uses a vocal prosthesis and

---

155 Daleks are intelligent extraterrestrials from the Doctor Who television series. They are violent and ruthless and bent on conquest. They seem to lack any ‘universal’ altruistic motivations.
156 The fiduciary non-cognitivist will include in this category some of the ‘amoralists’ that figure in debates about judgement internalism.
157 I hope to leave open the possibility that we can judge their actions wrong if we are thinking in terms of act-types; such judgements would involve preferences about what other moralists should do, not about what the creature in question should do.
A mistakes him for a Dalek.) It obviously would be a problem if fiduciary non-cognitivism predicted that, as a result, A would not extend her trust and commitments to B, and so would not disagree with B about the permissibility of abortion. The first objection to fiduciary non-cognitivism’s account of moral universality is that it cannot plausibly explain moral disagreement between two people one of whom believes that the other does not meet the conditions for being a moralist.

This objection only succeeds if I take a certain view about the nature of the generic trust attitudes that moralists must extend to other moralists. I’ll explain what this view is, and why I have good reason to reject it. I wish to take the view that the trust between moralists as such is not trust that moralists extend to those whom they believe meet the criteria specified by the terms of the moral trust network. Nor is it a more liberal version of this sort of trust, according to which moralists presumptively trust everyone whom they believe might meet the criteria, until they learn better. Rather, the trust between moralists as such is extended toward all and only those who actually do meet the criteria, regardless of the trusting parties’ contingent beliefs about who meets the criteria. This being so, A’s false beliefs about B are irrelevant to the question of whether or not they disagree about abortion.

There is nothing puzzling about the idea that someone might trust someone else without realizing it, and while having beliefs about that person (picked out in a certain way) that do not incline her to trust him. Superman and Clark Kent are the same person. But Lois can trust Superman to rescue a certain man from the top of a burning skyscraper, while not trusting Clark Kent to do so, if she does not realize that they are the same person. Similarly, a person may have generic trust in firefighters (or ones who meet certain criteria, e.g., being non-psychopathic) to rescue people from burning buildings that are near them, while failing to trust the person she sees next to her to rescue someone from a nearby burning building, because she is unaware that he is a firefighter.158

It is true that such a person may properly say ‘I didn’t trust him to rescue the man, because I didn’t realize that he was a firefighter.’ And yet it also seems correct to say that she did trust the man who he really is to perform the rescue, because she trusts firefighters to rescue people in his situation, and he is a firefighter. We don’t take the

---

158 I include this similar example to show that the special relation of numerical identity is not important here, as it is for many cases involving Superman in philosophy of mind and philosophy of language.
fact that Lois, upon learning that Clark Kent and Superman are the same person, can properly say, ‘I didn’t trust Clark to rescue the man because I didn’t realize he was Superman’ to imply that she did not trust the man who Clark Kent really is to perform the rescue.

We might say that Lois’s failure to trust Clark Kent before she learned his true properties is a failure of subjective trust, which we may say is a kind of trust that is relativized to an agent’s contingent beliefs. Objective trust, by contrast, is trust relative to the objective facts. Subjective trust is, I take it, a mental state with the behavioural and phenomenological symptoms that we typically associate with trust; for instance, if A subjectively trusts B to φ, she will feel disappointed or betrayed if she thinks that B did not φ, and might chastise B for the perceived failure. Objective trust does not in itself have any behavioural or phenomenological symptoms, although it does involve dispositions to form or lose subjective trust attitudes, depending on the trustor’s beliefs about the identity of the objectively trusted party.

Objective trust is, I think, what is important to an explanation of robust non-cognitive disagreement. Suppose that John and his mother have the preferences they report in Football (Mother and Son), and that neither has yielded or waived. If John (and only John) is chatting with the lady from the Gas Company, and John’s mother doesn’t realize that ‘the boy who is chatting with the lady from the Gas Company’ is John, John’s mother nonetheless disagrees with the boy who is chatting with the lady from the Gas Company about whether he is to play football. (Although we cannot use a ‘conciliation test’ to determine the robustness of her disagreement with this boy, it is obvious that her disagreement with him is of the same kind as her disagreement with John in the original case.) Subjective trust is not among the conditions for robust non-cognitive disagreement, though it is relevant to explaining why a person does or does not take herself to be a party to what I am calling a robust non-cognitive disagreement.

I shall retroactively stipulate that the entrusted concern relationships that figure in the Entrusted Concerns Hypothesis are relationships of objective trust.\(^{159}\) The moral

\(^{159}\) I don’t think I need to stipulate that entrusted parties who participate in robust non-cognitive disagreements must have ‘objective commitments’ to live up to their partners’ trust. An ‘objective commitment’ would be a commitment that does not depend on the committer’s beliefs. We do not seem to need the notion of an objective commitment, because there doesn’t seem to be such a thing as ‘subjective commitment’. Lois can say, ‘I didn’t trust Clark because I didn’t realize he was Superman.’ But a person cannot properly say, ‘I didn’t have a commitment
trust network posited by fiduciary non-cognitivism is a network of ‘objective’ entrusted concern relationships. Both the trust and the commitments of moralists as such are extended to everyone who in fact meets the criteria specified by the terms of the moral trust network, rather than to those whom they believe or presume meet those criteria.

What this means is that, according to the fiduciary non-cognitivist, the convinced Nazi who (atypically, I imagine) believes that Jews are literally subhuman vermin will still have moral disagreements with Jews, because his belief is false. Likewise the Seventeenth-Century Iberian who buys slaves from Africa so that he can work them to death in his South American mines, and who regards these slaves merely as cattle, will probably have moral disagreements with his slaves. Or at least, the false beliefs of the Nazi and the slave-owner do not stand in the way of disagreement. There are certainly other reasons to worry that the parties will lack the trust relationships required for disagreement in voiceable preference; I discuss one of these reasons in 5.4.3.

5.4.2 Moralists’ Trust is ImplausiblyInsensitive to Trustworthiness

To account for moral universality, the fiduciary non-cognitivist is forced to postulate that the threshold of trustworthiness for membership in the moral trust network is extremely low, and that the trust of moralists as such is surprisingly indiscriminate. The second objection to fiduciary non-cognitivism’s account of moral universality is simply that this postulation is highly implausible.

Since we can have moral disagreements with inveterate hypocrites (at least if they are human, non-psychopathic, non-feral, and so on) and the persistently ‘weak-willed’, and disapprove of their token actions, it must be the case that being highly unscrupulous does not make one an outcast from the moral trust network. Everyone must (tacitly etc.) trust the hypocrite, to some degree. Trusting the highly akratic in this way doesn’t seem particularly wise, and it is prima facie implausible to posit that we all do it. To support this postulation, I would need to defend some further evolutionary hypotheses. I need it plausibly to be the case that a moral trust network with very permissive membership criteria would have served the interests of early moralists. This would explain, given my earlier conjectures, why we would have evolved dispositions to trust that were highly indiscriminate with respect to trustworthiness. Properly defending the relevant hypotheses would of course involve showing that they fit the evidence. But as before, I
to Clark yesterday, because I made my promise to Superman, and I didn’t know then that Clark was Superman.'
shall at this point merely content myself with stating the hypotheses and showing that they do not defy common sense.

Imagine again the situation of our ancestors prior to the evolution of our innate and unconscious disposition to form moral entrusted concern relationships. I have conjectured that members of early moral trust networks would have induced people to join and stay in the network by making life unpleasant for outsiders. Because of this, I have supposed, it is very plausible that most people in this situation would have wanted to have been members. It is also somewhat plausible that people would have been willing to have been members even if they were not willing to do much of what membership required of them. For one thing, if they were members they would not have been the object of ill-treatment due to the mere fact that they were outsiders; in the normal course of events moralists actually have to be suspected of either unscrupulousness or perceived moral wrongdoing before being the object of ill-treatment at the hands of moralists as such. (By ‘perceived moral wrongdoing’ I mean doing what people judge to be morally wrong, whether or not this involves failing to live up to one’s obligations qua party to moral entrusted concern relationships.) For another, the ill-treatment to which they were subject would normally have taken the form of ‘blame’ and ‘punishment’ behaviours rather than simple hostility. These behaviours are a special focus of moral evaluation, and are, in commonsense morality, tightly constrained by moral norms. The costs they impose on ‘offenders’ are likely to be more moderate, more predictable, and more sensitive to what the agent had control over, than the hostile treatment inflicted on outsiders who perform altruistic-frustrating actions. They are also more responsive to the behaviour of perceived offenders after they have offended. According to commonsense morality, the propriety of punishment of a given degree of severity normally depends somewhat on whether the offender apologizes and exhibits contrition. Being a party to the network potentially gives a person the option of moderating her ill-treatment by repenting (this requires that

---

160 This raises a somewhat tricky question about why the institution of morality would not itself have inhibited moralists from making life especially unpleasant for outsiders, since making life unpleasant for people is intrinsically frustrating to altruistic desires. I will assume, but won’t defend, the following answer to this question. The moral judgements of early moralists tended to prescribe or permit treating outsiders worse than moralists, because they tended to prescribe or permit actions that ‘improved’ people’s behaviour (i.e. made it more satisfactory to altruistic desires). Early moralists believed that treating outsiders relatively badly was likely to ‘improve’ their behaviour. Such treatment gave the outsiders an incentive to become moralists, and moralists tend to behave ‘better’ than non-moralists because, since they have committed to do as their moral trust relationships require, their reputations will be harmed if they do not.
she judges that she did wrong, which is only possible if she is a network member). So it would have been in the interest of those who are not willing to be scrupulous to be members of the network.

I hypothesize that it would also have been in the interest of moralists that people who are not willing to be scrupulous should be members of the network (as long as those people meet certain minimal criteria, which include a minimal willingness to participate scrupulously conditional on being members). If someone became a member her behaviour was likely to ‘improve’ somewhat, in the sense that it would frustrate altruistic desires less, because her reputation was on the line. People wouldn’t trust her so well in other matters if she didn’t live up to her commitment to act only according to those rules conformity to which would best satisfy her altruistic desires.

I also find it plausible to suppose that the policing and punishment of unscrupulous moralists tended to be either less effortful for other moralists, or more satisfying to moralists’ altruistic desires, than measures taken to deter altruistic-desire-frustrating actions of outsiders to the network. The machinery of punishment, blame, forgiveness, repentance, and so on, facilitated by the moral trust network, provided a means of controlling altruistic-desire-frustrating behaviour that was peculiarly satisfying to altruistic desires associated with desert. Moralists—including unscrupulous ones—are obliged to honour other moralists’ discursive entitlements, which means being somewhat forthcoming about their reasons for acting. This meant that moralists were likely to have more information for a given expenditure of effort about what other moralists did and why, which would have helped to achieve a tighter correlation between the imposition of costs and the actual performance of altruistic-desire-frustrating actions. This would have been more satisfying to altruistic desires, and also would have lowered the costs of the cost-imposition itself, reducing both wasted effort and the social conflict that cost-imposition will sometimes produce. (I admit that this is a bit hand-wavey; I need to do more work to spell out the causal links I am positing here, and to explain the dependence of the ‘machinery’ of blame etc. on the institution of morality.)

I have said that a person puts her reputation on the line when she becomes a moralist, and that people are induced to become moralists by the threat of ill-treatment. It might be objected that these two claims are in tension. Humans—modern ones anyway—do not in general think less of a human if she fails to live up to a commitment that she was coerced into making, so the reputation of a person who becomes a moralist
to avoid ill-treatment cannot really be at stake. This is plausible if the mere threat of violence were all that induced potential moralists to make the commitment. But I think people would normally have had somewhat more complex motivations for joining the network than this—including perhaps desires to be part of society, desires for esteem, and desires not to be resented as freeriders (i.e. for failing to reciprocate others’ forbearance from frustrating altruistic desires, or for failing to commit to do so). If fear of violence and the like figured prominently rather than exclusively among people’s motivations for committing to join, alongside more ‘pro-social’ motivations of the sort I have mentioned, it seems plausible that they would have genuinely put some part of their reputation at stake by committing to live up to other moralists’ trust.

So it is in the interest of moralists to let people with poor motivations to be scrupulous into the network, and in the interest of those people to join the network. This explains why our evolved dispositions to trust and commit as membership in the network requires would be dispositions to do so in a way that was highly indiscriminate with respect to trustworthiness and willingness to participate scrupulously.

5.4.3 Trust Between Antagonistic Agents

Surely if the animosity between the parties in the blackmail case I considered in 1.2.4 made it implausible that the parties shared an entrusted concern relationship, the antagonistic relationship between Nazis and Jews in Auschwitz ought to make it implausible that they entrusted to each other concerns to accommodate each other’s altruistic desires. The fiduciary non-cognitivist must posit such trust relationships if she is to explain how they could be in moral disagreement, but this postulation is intuitively implausible. This is the third objection to fiduciary non-cognitivism’s explanation for moral universality.\(^{161}\) There is only so much I can say in response to the objection, which highlights a real and serious cost for the theory. However, I can at least point out two important differences between the blackmail case and the case of the mutually trusting Jew and Nazi.

Two agents can be hostile toward each other, in a way that would prevent any personal relationship of trust from forming between them, and yet still trust each other in certain ways. Two soldiers A and B in the same army might utterly despise each other, to such an extent that each would never do the other the smallest personal favour if asked, nor ever miss a convenient opportunity to belittle or humiliate the other. Yet A

\(^{161}\) Lachlan Umbers has pressed this objection against my theory.
might nonetheless to some extent trust B to help A when B’s duty as a soldier required
B to (and vice versa). Here A trusts B qua occupant of a certain role, to help A qua
occupant of a certain role. I propose that the trust between a Jewish prisoner and a Nazi
commandant in Auschwitz that allows them to be parties to moral disagreements is
analogous. They share relationships of trust and commitment qua moralists, even if they
do not share them in any personal capacity. There is no way to plausibly interpret the
parties in the blackmail case as sharing a trust relationship in some non-personal
capacity that would make their reported preferences in Restaurant (Blackmail II)
voiceable, and the fiduciary non-cognitivist does not posit anything that would suggest
otherwise, so the theory doesn’t seem to predict more implausible trust between
antagonistic agents than it must in order to explain the metaethical data.

Second, given the fiduciary non-cognitivist’s wider claims, the posited trust
relationship between Jew and Nazi is in one respect more plausible than the postulation
of a non-moral trust relationship between the parties in the blackmail case would be.
According to the fiduciary non-cognitivist, neither the Jew nor the Nazi risks being
obliged to make any significant special sacrifice by extending her trust and
commitments to the other as a consequence. To be sure, each takes on some weak
obligation to honour the other’s discursive entitlements. But outside of conversations,
neither is obliged to do anything for the other that she is not also obliged to do for every
other moralist. By contrast, an entrusted concern relationships between the parties in the
blackmail case, of a sort that would make their reported preferences voiceable, would
create pressure on the parties to go out of their way to accommodate each other’s
weighted desires. This makes mutual trust between Jew and Nazi less psychologically
implausible, since it is implausible to suppose that either party would be willing to do
the other any special favours. (Similar considerations also probably account for some of
the plausibility of the idea that there is trust in the soldier case: other soldiers, not just
A, trust B to help A when B’s military duty requires it, and B will be concerned to live
up their trust.)

5.4.4 Moral Judgement Without Altruistic Desires?
Fiduciary non-cognitivism almost certainly predicts that agents that wholly lack
altruistic desires cannot be parties to the moral trust network. It is possible (depending
on what we say about ‘simplicity desires’) that a hypothetical moralist without altruistic
desires simply could not form discretion-licencing preferences; if so, she couldn’t help
but violate some moralist’s trust in almost any decision she faced. If the moralist could form discretion-licencing preferences based on her simplicity desire alone, her conformity to those preferences would have little chance of satisfying other moralists’ altruistic desires. The moral trust network exists as a means to improve the satisfaction of altruistic desires, and her inclusion in the network couldn’t be expected to further this end. So we should expect the terms of the moral trust relationship to exclude agents without altruistic desires from the set of agents whom moralists as such are required to trust. It is indeed somewhat plausible that the criteria would be more restrictive than this, excluding those who lacked a certain range of altruistic desires, or whose altruistic desires were not too vastly different in their relative strengths from those of ordinary human beings.

This leads to our fourth objection, which is that there are cases of what seems to be moral disagreement in which one of the parties plausibly lacks altruistic desires. I shall consider a few candidate cases in a moment. Before doing so, let me remind the reader that, as I am using it, ‘altruistic desire’ is a technical term, whose precise meaning I have left open. It encompasses some desires that are not clearly altruistic in the ordinary sense, such as desires that people should get what they deserve, even if they deserve suffering, and permission desires, which may be satisfied when a person who would rather not help others forbears from helping them. The term could perhaps be stretched to accommodate desires connected with aesthetic disgust or with ideas of hygiene and pollution—this might be needed to account for apparently moral judgements forbidding such things as masturbation. The desires that I am calling ‘altruistic’ must meet these two criteria however: they must be ‘universal’, and they must be fairly common among human beings.

Altruistic desires are, I take it, adaptive, and susceptible to social, cultural, and environmental influences. For instance, in societies whose harsh environments make providing for non-productive members impossible, we should not be too surprised to find that people’s altruistic desires are frustrated to a lesser degree by the killing of disabled infants and the casting out of old and infirm members. The fiduciary non-cognitivist will certainly wish to posit that this sort of variation, even in quite extreme cases, will not exclude agents from the moral trust network. This postulation doesn’t seem initially implausible. Even an agent whose altruistic desires differ quite significantly from the norm in their relative strengths will usually satisfy other moralists’ altruistic desires better by acting on her altruistic desires than she would by
acting on her partial and self-interested desires, and the influence of the moral trust network on members is, broadly speaking, to encourage them to give greater weight to their altruistic desires in their decisions. So we should expect the criteria for membership in the moral trust network to be permissive with regard to this sort of adaptive variation, even in fairly extreme cases.

Thus I think that fiduciary non-cognitivism quite plausibly predicts that there would be moral disagreement between the missionary and the cannibals in R. M. Hare’s famous case.\(^\text{162}\) Admittedly, I can’t claim that it predicts that the missionary and the cannibals will be in moral disagreement on the question Hare describes, viz., whether a good man is one who is meek and mild, or one who is bold and burly and collects plenty of scalps. For one thing, I have not offered a fiduciary non-cognitivist account of goodness judgements. For another, I am sceptical of whether the cannibals’ judgement could be a judgement of moral goodness.\(^\text{163}\) However, there are sure to be moral disagreements of a narrowly normative sort in the vicinity. For instance, the cannibals judge that morality permits you to kill innocent people and collect their scalps, and the missionary judges that morality forbids this. There is no reason why the cannibals’ permission judgement could not be supported by their altruistic desires; they need merely have weak anti-homicide desires, and certain strong permission desires, which are satisfied when agents who wish to pursue the non-moral glory of killing and scalping do so.

Hare’s description of the case gives us little reason to think that the altruistic desires of the cannibals differ vastly from the norm. Presumably in many matters the cannibals’ moral judgements will be reflective of fairly normal altruistic desires. They will probably judge that lying, cheating, and promise-breaking are wrong; they will disapprove of some theft; and their tolerance of killing for glory will probably be limited in certain ways (e.g. one should not kill those to whom one owes loyalty for their scalps). In short, Hare’s cannibals, if they existed, would presumably have been

---


\(^{163}\) Moral excellence plausibly cannot depend on one’s talent or physical abilities—things over which one lacks control—in the way that non-moral excellence can (Susan Wolf, ‘Moral Saints’, *Journal of Philosophy* 79, no. 8 (1982): 434). Being such that one collects more scalps than the average is presumably judged to be a non-moral excellence, because people are limited in their ability to collect scalps by their natural abilities and talents. Likewise, being burly is presumably judged to be a non-moral excellence, because it is not in everyone’s power to be burly.
welcome members of the moral trust network, and so capable of being in moral disagreement with Hare’s missionaries.

Can we imagine creatures that obviously lack altruistic desires making moral judgements or having attitudes of moral disapproval? A race of thoroughly sadistic aliens might disapprove of all altruism, and judge that altruism should be eschewed, but I find it hard to imagine that their disapproval might be moral disapproval, or that their normative judgement might be a judgement that all altruism is morally wrong. It does seem to me that their normative judgements might put them in disagreement with humans, and I am not sure how to explain this. But since the aliens’ judgements are apparently not moral judgements, this presumably cannot be a moral disagreement. Accounting for it is thus beyond the scope of this work.

**Conclusion**

In this chapter I have made a case for the existence of the moral trust network. I explained in a very rough and conjectural fashion the circumstances in which it could have come to exist, and how humans might have the evolved dispositions that they would have needed to have had for it to have persisted without anyone possessing more than tacit knowledge of it. I have argued that there are non-moral examples of trust networks that share features of the moral trust network that we might otherwise have thought exotic—that there are trust networks whose members are mostly strangers to each other, for instance. And I have argued that, for cognitivists and non-cognitivists alike, providing a wholly satisfactory explanation of moral approval and disapproval requires postulating the moral trust network or something similarly immodest.

---

164 The same is true *a fortiori* of other agents who, though evil, surely would have had altruistic desires that were more similar to our own than the cannibals’, such as Hitler and others who held high office in the Third Reich. Nazis’ moral judgements about matters unrelated to race, eugenics, genocide, and the role of the state in society, were not particularly distinctive. And their most distinctive moral judgements were in part explained by peculiar descriptive beliefs about race and biology.
Moral Rules

In this chapter I present an incomplete explanation for the apparent rule-based character of moral requirements and the value of simplicity in moral theorizing. These characteristics of moral thought and talk are, I propose, explained by the advantages for moralists of making it difficult to abuse trust without detection.

Section 6.1 describes a method of abuse-detection. Parties to entrusted concern relationships can detect abuses of trust that are concealed by insincere voicing of preferences, by comparing such preferences to preferences voiced by the same party concerning similar situations in the past. All things being equal, it would serve moralists’ interests for the comparative method of abuse detection to be as feasible as possible. In 6.2 I propose that the comparative method of abuse detection is more feasible if (a) moralists’ voiceable preferences are preferences for conformity to norms rather than the performance of particular actions, and (b) moralists are somehow induced to prefer that they and others should follow simpler norms. This influence, I propose, is brought about by, at least in effect, including ‘simplicity desires’ (desires for conformity to simpler norms) among moralists’ weighted desires. Section 6.3 refines (a): it clarifies the nature of the norm-conformity decisions in which I am proposing that moralists have voices.

In 6.4 I posit that moralists fundamentally have voices in one another’s choices of which comprehensive sets of norms to follow. They have voices in choices of which non-comprehensive sets of norms to follow only derivatively. I discuss various reasons why moralists would want to voice preferences for conformity to non-comprehensive norms. In 6.5 I give my long-postponed explanation of why bringing it about that moralists (in effect) have weighted simplicity desires would better facilitate the comparative method of abuse-detection. My explanation for the apparent virtuousness
of simplicity in moral reasoning is incomplete because I have, as of writing, failed to come up with a good account of why moralists would have simplicity desires, weighted or otherwise, or of how the terms of the moral trust network could be engineered so as to bring it about that they have such desires. Section 6.6 discusses the desiderata such an explanation must meet.

### Section 6.1 The Comparability Desideratum

One of the principal influences on the terms of the moral trust network will be the need to make abuses—chiefly actions done in defiance of discretion-licencing preferences—as easy to detect as possible. Moralists, by exercising their discursive entitlements, can exert pressure on other moralists to voice preferences, so that their future deeds can be measured against their words. But a moralist may voice preferences insincerely, falsely presenting ‘extra-discretionary’ actions as licenced by her voiceable preferences, and so avoid discovery. In this brief section I shall describe a simple method for detecting abuses and comment on the prospects for its use by moralists. In the sections that follow I argue that the rule-based character of moral prescriptions and the importance placed on simplicity in moral theorizing are largely explained by the desirability, from moralists’ perspectives, of facilitating the use of this method.

In entrusted concern relationships generally, it is sometimes possible to detect whether a voiced preference is voiceable by comparing that preference to preferences voiced by the same entrusting party in similar situations in the past. If John’s mother voiced a preference that John should play football, then made a bet with someone that John would go the year without breaking any bones, and subsequently voiced a preference that John should not play football, it would be reasonable to infer that her newly voiced preference was not genuinely accommodation-obliging. At least, this would be a reasonable inference if the contexts in which she voiced her differing preferences were in other respects very similar. For then we could infer that John’s mother’s new preference was explained either by a change in her weighted desires, or a change in her beliefs about the pecuniary advantages for her of John’s remaining uninjured; and, since changes in basic desires are far less common than changes in derived desires resulting from changes in beliefs, the latter is a much more likely explanation. Roughly the same method can sometimes be used to determine whether an entrusted party’s decision was made in conformity to his discretion-licencing
preferences: the decision can be compared to decisions he has made, and preferences he has voiced qua entrusted party, in similar situations in the past.

It would clearly serve the interests of moralists for them to have opportunities to make these sorts of abuse-revealing comparisons to the greatest possible extent, all else being equal. And, as members of a dual-order fully connected network, whose weighted desires, for all I’ve said so far, consist solely of altruistic desires, members of the moral trust network have an advantage in this regard over parties to entrusted concern relationships that do not belong to larger networks. Moralists can fruitfully compare preferences voiced and actions performed by a particular agent with respect to similar sorts of decisions, regardless of the identities of the agent making and the patients affected by those decisions, and regardless of whether that member voiced or performed those preferences or actions qua first-order entrusted party (and the agent in one or both of the decisions), first-order entrusting party, second-order entrusted party, or second-order entrusting party. This is because (i) her weighted desires (as I am presently supposing) consist of all and only her altruistic desires, and hence are universal; and (ii) the terms of her moral trust relationships guarantee harmony between her preferences qua first-order party and qua second-order party.

As a result:

For all moralists S, and
all F such that F is a complete set of descriptive properties of an action, and
all times t₁, and
all moralists B₁, and
all actions φ such that φ is F and φ is available to B₁, and
all moralists A₁ such that A₁ prefers qua first-order entrusting party that B₁ should not-φ and A₁ has not waived his demand that B₁ should not-φ;
if there is some moralist B₂, and
some time t₂
some action ψ such that ψ is F and ψ is available to B₂, and
some moralist A₂ such that A₂ prefers qua first-order entrusting party that B₂ should not-ψ and A₂ has not waived his demand that B₂ should not-ψ, and
S is rational, and S’s weighted desires have not changed between t₁ and t₂;
then ([(S is B₂ and S prefers at t₂, qua first-order entrusted party, that B₂ should ψ), or
S prefers at t₂, qua first-order entrusting party, that B₂ should ψ, or
S prefers at t₂, qua second-order entrusted party, not to waive her demand that B₂ should ψ, or
S prefers at t₂, qua second-order entrusting party, that A₂ should waive her demand that B₂ should not-ψ]
iff [(if $S$ is $B_1$, then $S$ prefers at $t_1$, qua first-order entrusted party, that $B_1$ should $\varphi$), and $S$ prefers at $t_1$, qua first-order entrusting party, that $B_1$ should $\varphi$, and $S$ prefers at $t_1$, qua second-order entrusted party, not to waive her demand that $B_1$ should $\varphi$, and $S$ prefers at $t_1$, qua second-order entrusting party, that $A_1$ should waive her demand that $B_1$ should not-$\varphi$].

If a moralist voices or purports to act on preferences that she could not have if this biconditional were true, this is evidence that she is irrational, or that her basic desires have changed between $t_1$ and $t_2$, or that the preferences she has voiced or purported to act on are not her true voiceable preferences. Sometimes this last possibility will be the most likely. So if a moralist doesn’t practice what she preaches, or if she makes different demands of different people in the same circumstances, she potentially betrays unscrupulousness. Parties to more intimate and personal entrusted concern relationships, which are not part of larger trust networks, will not be able to glean evidence of unscrupulousness from such a wide variety of discordances among ostensible preferences.

On the other hand, parties to such intimate and personal entrusted concern relationships will not need so many ways to compare their partners’ ostensibly voiceable preferences for evidence of abuse, because each party will know the other well and know her wants. And here we come to a major disadvantage of the moral trust network: it is, as I have said, a network of strangers. For any given moralist, there will often only be a small number of other moralists around who have had the opportunity to observe her actions and demands over time, and there may be no one around who knows her intimately. So those who observe her will frequently have no idea whether the preferences she voices or purports to act on are in tension with those she has voiced or purported to act on in the past. This problem is exacerbated by a general problem for the comparative method, which is that no two situations are exactly alike. Even at the best of times, a party to an entrusted concern relationship can usually find some consideration to cite that might explain an apparent discrepancy in the history of her actions and voiced preferences.

6.2. ‘Norms’ and Simplicity

One way in which to better facilitate the comparative method of abuse-detection would be to place some kind of constraint on the potential diversity of a moralist’s voiceable
preferences across all options in all contexts. I propose that this explains our evident disposition to treat simplicity as a virtue in moral theorizing. We are disposed to prefer simpler theories about right and wrong to complex ones, and the knowledge that we have this disposition makes it easier for others to infer from our public judgements in some cases what our judgements in different cases are likely to be, if we have reasoned well and judged sincerely. The terms of moral trust relationships have been shaped in such a way as to give us this disposition, because it serves the purpose for which the moral trust network was created for our abuses to be easily detectible via the comparative method. It will be a while before I can make clear what moral theorizing is according to fiduciary non-cognitivism, but I can make progress toward explaining our disposition to prefer simple theories by explaining how considerations of simplicity affect moralists’ voiceable preferences. Once I have done this, I will give a more careful explanation of the advantages of simplicity.

Very roughly, as I anticipated in Chapter 4, I am going to propose that moralists have voices in other moralists’ decisions concerning which norms they are to conform to, rather than in their decision concerning which particular actions to perform. And I am going to propose that moralists will tend to have voiceable preferences that other moralists should conform to simpler norms, because their voiceable preferences are influenced by ‘simplicity desires’. Before I can flesh out these ideas, I must introduce some new terms, the most important being ‘fundamental entitlements’, and ‘comprehensive norm-partitions’, and make some hypotheses about the decisions in which parties to entrusted concern relationships have voices.

6.2.1 Fundamental and Derived Entitlements

Before I can address them I need to make clearer what in general is involved in preferring one option to another, on my usage of these terms. When I formulated the Entrusted Concerns Hypothesis in 1.2.3, I said that, to be in robust non-cognitive disagreement with one another, parties to an entrusted concern relationship had to have, and merely had to have, preferences (based on the right kind of desires, and, as I would now add, on the right kind of doxastic attitudes) for mutually incompatible, jointly exhaustive options (in those decisions of the entrusted party in which the entrusting party had a voice). On this formulation, the options concerning which the parties to a robust non-cognitive disagreement have disagreeing preferences are extremely coarse-grained. In most cases these options will have numerous ‘sub-options’—different ways
in which the entrusted party can realize them. For instance, the option *John plays football* may be realized by the sub-options *John plays goal-keeper* or *John plays offence*, assuming that it is up to John which of these positions he plays. I will call an option or sub-option that does not itself have any sub-options—i.e. that is not multiply realizable—an *atomic* option or sub-option.

Parties to robust non-cognitive disagreements will typically find some sub-options preferable to others. It is possible, and indeed would be common, for some pair of incompatible options $\phi$ and $\psi$, for a party to both prefer some sub-options of $\phi$ to some sub-options of $\psi$, and vice versa. So let me clarify that, when I say that a person $S$ prefers $\phi$ to $\psi$, where these are mutually incompatible and multiply realizable options, I mean that out of all the atomic sub-options of $\phi$ and $\psi$, the atomic sub-option $S$ best prefers is a sub-option of $\phi$. Person $S$ prefers option $\phi$ to option $\psi$, where $\phi$ is multiply realizable but $\psi$ is not, if and only if there is some atomic sub-option of $\phi$ that $S$ prefers to $\psi$. This of course applies to voiceable preferences as well as all-things-considered preferences.

I have allowed that entrusted concern relationships may give parties voices in some decisions and not others, and this ought to constrain the range of sub-options whose weighted utility (i.e. utility understood as weighted desire satisfaction) could potentially influence a voiceable preference for the options they realize. I will call an option (or sub-option) in whose a selection a party to an entrusted concern relationship has a voice a *live* option, or an option that is *live for that party*. Suppose that a party $S$ to an entrusted concern relationship has a voice, as such, in a decision between options $\phi$ and not-$\phi$, and that there is a decision $d$ such that (i) $S$ does not have a voice in $d$, and (ii) there is some option $\alpha$ such that $\alpha$ is available in $d$ and $\alpha$ is an atomic sub-option of $\phi$. Since $\alpha$ is not a live sub-option for $S$, I propose that the weighted utility of $\alpha$ cannot affect $S$’s voiceable preference between $\phi$ and not-$\phi$. If (i) $\beta$ is a live atomic sub-option of not-$\phi$, and (ii), out of all the atomic sub-options of $\phi$ and not-$\phi$, $\alpha$ has the highest weighted utility for $S$ and $\beta$ the unique second-highest, then $S$ has a voiceable preference for not-$\phi$ over $\phi$. It is thus the weighted utility of live atomic sub-options that determines a party’s voiceable preferences.

I will now propose a distinction among entitlements to voices in decisions, a distinction between ‘fundamental’ and ‘derived’ entitlements. A party’s entitlement to a voice in a decision $d$ is *fundamental*, in the ‘scope’ sense with which I am now concerned, if and only if $d$ is a choice between live atomic sub-options. For all options $\phi$
and ψ, a party S has a derived entitlement to a voice (or simply a derived voice) in a choice between φ and ψ if and only if (i) at least one of these options is multiply realizable; (ii) φ and ψ are mutually exclusive; (iii) there is some α such that α is an atomic option, and α is live for S, and either α is φ or α is a sub-option of φ; and (iv) there is some β such that β is an atomic option, and β is live for S, and either β is ψ or β is a sub-option of ψ. On this view, a party’s entitlements to voices in decisions supervene on, and are explained by, her entitlements to voices in decisions between atomic sub-options. Thus to specify which decisions a party has a voice in, the terms of an entrusted concern relationship need only specify which atomic sub-options are live for her.

I shall now hypothesize that this (and this alone) is how the terms of entrusted concern relationships do specify the decisions in which parties have voices. This makes for a parsimonious theory of entrusted concern relationships. We need to postulate that parties have voices in choices among some atomic sub-options and not others in order to explain why a moralist might fail to have a voiceable preference for φ even if the atomic sub-option with the highest weighted utility for her is a sub-option of φ. Once we postulate this, we can explain parties’ voices in choices among non-atomic options as deriving from their voices in choices among atomic sub-options. We have no apparent need to postulate that entrusted concern relationships entitle parties to voices in choices among non-atomic sub-options via any other means. So simplicity favours the view that they never do.

Fundamental entitlements explain derived entitlements. An entrusted party is obliged to accommodate a preference that he should φ, where his partner has voiced this preference in exercise of a derived entitlement, only because he is obliged to accommodate the preferences she voices in exercise of some fundamental entitlements. Likewise, a preference voiced by an entrusted party in exercise of a derived entitlement to a voice in a decision is only discretion-licensing because it is explained by his possession of discretion-licencing preferences that are voiceable in virtue of fundamental entitlements.

6.2.2 ‘Norm-Partitions’ and Related Terms

The following discussion of norm-partitions contains some repetition from 4.1.1. A norm-partition is a pair of an extension—a set of ‘action-circumstance pairs’—and a prescription function—a function from each action-circumstance pair in that set to a
prescription to perform the action in that circumstance or not perform it.\textsuperscript{165} An action-circumstance pair is a pair of an action and a possible circumstance in which it is performed. An action-circumstance pair is ‘maximally specified’: action-circumstance pair \(\langle \phi_1, C_1 \rangle\) and action-circumstance pair \(\langle \phi_2, C_2 \rangle\) are numerically identical just in case every detail about the actions, and about the circumstances in which they occur, is exactly similar. The action and circumstance in an action-circumstance pair must be such that the agent is, in the circumstance, able to identify the action and has a choice to perform or not perform it.

Norm-partitions make prescriptions. Any action that the prescription-function of a norm-partition requires (in its relevant circumstance) is one that the norm-partition requires, and any it forbids is one that the norm-partition forbids. A norm-partition cannot require two different actions in the same circumstance—this is a constraint on possible prescription functions. A norm-partition also cannot forbid all the actions that are available in a single circumstance. Thus, for all \(\langle \phi, C \rangle\) such that \(\langle \phi, C \rangle\) is an action-circumstance pair, a norm-partition requires \(\phi\)-ing in \(C\) if and only if it forbids not-\(\phi\)-ing in \(C\).

Sometimes we may identify the prescriptions of a norm-partition in a circumstance or range of circumstances \(C\) by a property that is correlated with those prescriptions in \(C\). For instance, we might say with respect to \(C\) that, according to norm-partition \(N\), stealing is forbidden, meaning that \(N\) forbids all acts of stealing in \(C\). For all properties \(F\), \(N\) forbids (requires) \(F\) actions \textit{necessarily} if and only if \(N\) forbids (requires) all possible \(F\) actions; if \(N\) forbids (requires) \(F\) actions in \(C\), but does not forbid (require) \(F\) actions necessarily, then \(N\) forbids (requires) \(F\) actions in \(C\) \textit{contingently}.

A \textit{norm} is a norm-partition forbidding or requiring actions with a certain set of properties. When I talk of ‘norms’ I usually mean the complete set of prescriptions made by a relevant norm-partition. An agent \textit{follows} a norm-partition, in a given moment or over a period of time, if she performs no action that it does not require of her in that moment or over that period. Thus she fails to follow the norm-partition if she either performs an action it forbids (or ‘defies’ the norm-partition), or because she acts

\textsuperscript{165} ‘Norm-partition’ is admittedly an infelicitous label. The idea behind it is that the prescription function of a norm-partition ‘partitions’ the set of action-circumstance pairs in its extension into required or forbidden subsets. This is not quite the case, however, for the prescription function may output only requirements, or only prohibitions, in which case the extension is not partitioned.
in a circumstance that falls outside of its extension. A *comprehensive* norm-partition is one whose extension is the set of all possible circumstances of action.

Norm-partitions $N$ and $O$ make *incompatible prescriptions* in some circumstance $C$ just in case either $N$ requires an action in $C$ that $O$ forbids or $N$ forbids an action in $C$ that $O$ requires (otherwise they make *compatible prescriptions* in $C$). $N$ is *incompatible* with $O$ just in case there is some circumstance in which they make incompatible prescriptions (if there isn’t, they are *compatible*). $N$ *entails* or *implies* $O$ just in case the extension of $O$ is a subset of the extension of $N$, and $N$ and $O$ are compatible.

A norm-partition can be more or less simple. Defining simplicity is notoriously difficult, and I must rely in what follows on our intuitive grasp of the concept. But among norm-partitions of the same extension, one norm-partition is *typically* simpler than another if it is possible to state in full which actions it requires and which it forbids in fewer words using the same natural language.

6.2.3 A Restatement of the Proposal, and Two Desiderata Derived From the Comparability Desideratum

I propose that moralists, qua parties to the first-order moral trust network, have fundamental entitlements to voices in, and only in, other moralists’ choices among which *comprehensive* norm-partitions they will conform to. (I use the term ‘conform to’ in a loose and intuitive way here; I consider below how best to precisify it.) And I propose that (i) moralists have desires that the comprehensive norm-partitions that moralists conform to should be simpler rather than more complex, and (ii) these desires are ‘effectively’ weighted for moralists. That is, these simplicity desires are not weighted, strictly speaking, because they are not basic desires. But they are insensitive to ordinary changes in moralists’ contingent beliefs, and the desires they are characteristically derived from are weighted, so we can loosely speak of moralists as having weighted simplicity desires.

I want my theory of the terms of the moral trust network to predict the following: that any perfectly rational moralist will, at any given moment, voiceably prefer that all agents in all circumstances should conform to the same comprehensive norm-partition, where the complexity of this norm-partition would be diminished in proportion to the strength of her simplicity desire. In other words, I want to predict that two states of affairs hold, which I call ‘Simplicity’ (with a capital ‘S’) and Invariance. *Simplicity* obtains just in case rational moralists always have voiceable preferences for conformity
to simpler norms, and the simplicity of these norms tends to increase as in proportion to the strength of their simplicity desires. *Invariance* is the state of affairs that holds just in case, for all moralists A, either A is not perfectly rational, or there is some comprehensive norm-partition N such that, for all moralists B, A voiceably prefers that B should follow N (A and B may be the same person, in which case the voiceable preference is a discretion-licensing one).

Simultaneously predicting both of these states of affairs will turn out to be difficult. Yet Invariance must obtain, or else Simplicity will not facilitate easier abuse-detection via the comparative method—or not as well as it might have. The comparative method is most easily used if a rational moralist must have like voiceable preferences in like cases, regardless of the identities of the parties involved in the cases—that way an observed discrepancy among her voiced preferences in like cases will be evidence of abuse. If a perfectly rational party may have voiceable preferences for conformity to a comprehensive norm-partition prescribing a certain action in one case, and for conformity to a comprehensive norm-partition prescribing a different action in an exactly similar case, then a discrepancy in her voiced preferences in those cases won’t betray abuse, however simple those comprehensive norm-partitions may be.

Note that, absent the influence of a simplicity desire, Invariance will automatically obtain. Or at least, if (i) a moralist has a voice in other moralists’ choices among which comprehensive norm-partitions to conform to, (ii) we assume that ‘conform to’ means ‘follow’, and (iii) her voiceable preferences are based solely on her altruistic desires, then there will be some comprehensive norm-partition N such that she voiceably prefers that all moralists should follow N. This is because altruistic desires are universal. Whether or not an action satisfies them does not depend on the identities of the people and places involved, or their relationship to the desirer; it depends only on the action’s properties.

If two moralists are in radically different circumstances, and if we continue to assume that their weighted desires consist solely of altruistic desires, mightn’t this give a moralist A reason to voiceably prefer that they should follow different norms that are ‘tailored’ to their circumstances? I cannot see what such tailoring might involve. The very same comprehensive norm-partition will give prescriptions that maximize the satisfaction of A’s altruistic desires in every possible circumstance. Concerns of simplicity do not require that the norm-partition be compromised in such a way that
following it in some circumstances would not optimally satisfy A’s altruistic desires, while following it in others would. There is thus nothing to be gained by ‘tailoring’.

6.3 ‘Conforming to’ Norms

6.3.1 Three Precisifications Considered

In order to predict Simplicity, I am, as I have already said, hypothesizing that moralists have voices in other moralists’ choices among which comprehensive norm-partitions they will conform to, and that they prefer that other moralists should conform to simpler rather than more complex norms because their preferences are influenced by their simplicity desires, which are desires that moralists should conform to simpler norms. But, given my aims, what should ‘conforming to’ a norm-partition mean here? This will affect what I say about the demands moralists can make of one another, and about the content of the simplicity desire. Here are four possibilities: (1) conforming to a norm-partition N could mean following N; (2) it could mean following N for some period of time; (3) it could mean necessarily following N, on some suitable understanding of ‘necessarily’. I will evaluate these proposals in a moment. Before proceeding, I should warn the reader that I will sometimes say that a moralist prefers, or has a preference for, a norm, as a shorthand for saying that she has a preference for conforming to that norm, on whatever precisification of ‘conform to’ is relevant.

(1) ‘Conforming to N’ as following N. Suppose that moralist A has a voice in B’s choice among which comprehensive norm-partitions to follow, and that A’s relevant simplicity desire is a desire that B should follow simpler norms. The problem with this idea is that it does not have the result that A will voice preferences that B should follow simpler norms. Or, less ambiguously, the problem is that it does not have the result that the comprehensive norm-partition N such that A has a voiceable preference that B should conform to N is simple. This is because B is bound to follow simple norms when he follows N, even if N is extremely complex. The simplicity desire is thus too easily satisfied. Whatever B does, there will be some extremely simple comprehensive norm-partition that requires that action, and B will count as following that norm-partition. If B lights a cigarette, there is a very simple comprehensive norm-partition that prescribes maximizing expected cigarette-lighting in all circumstances, which B has succeeded in

---

166 This is partly to avoid complex locutions like ‘S has a preference for conformity to norms conformity to which would satisfy such-and-such desires.’
following. A’s simplicity desire will have little effect on the norms she best prefers that B should follow. A will probably just form a preference for a very complex comprehensive norm-partition prescribing actions that would maximally satisfy A’s altruistic desires in all circumstances. If B does follows this comprehensive norm-partition, he is also sure to follow some other extremely simple comprehensive norm-partition prescribing whatever action B performs. So B’s norm-conformity choice will satisfy A’s simplicity desire very well regardless of the complexity of the norms that she demands he should conform to.

(2) ‘Conforming to N’ as following N for a period of time. Now suppose that ‘conform to’ means follow over a certain period of time (call this period d), and hence that A’s simplicity desire is a desire that B should follow simpler norms over d. This would partly address the problem that the last proposal faced. This is because it is not necessarily true that whatever B does over d there will be some perfectly simple comprehensive norm-partition such that B counts as following it. If B does a lot of different things over d, and responds to circumstances with similar features in a lot of different ways, the simplest comprehensive norm-partition he follows over d may not be very close in simplicity to the simplest comprehensive norm-partition he could have followed. So A’s simplicity desire probably will affect the simplicity of the norms that she demands that he should follow over d. It will be frustrated by B’s conformity to norms whose requirements are sensitive to a lot of distinct properties of circumstances and actions.

Unfortunately, A’s simplicity desire will not always affect A’s preferences. For if A knows that B is only going to face the same sorts of choice situations over d (whether or not she knows what those choice situations are), she will be in much the same situation as we imagined her to be in when we were evaluating the view on which to ‘conform to’ norms simply meant to follow them at a given moment. A will simply form a preference for the norms N whose prescribed actions would maximally satisfy A’s altruistic desires in all circumstances. N would presumably be very complex, but since B is in the same circumstances throughout d, N will always contingently require him to do the same thing throughout d. Thus if B does follows N over d, he is also bound to follow some other extremely simple norm-partition throughout d prescribing whatever action N contingently prescribed throughout d. A’s simplicity desire will be perfectly satisfied. So if A knows that B will be in the same circumstances throughout d,
whether or not she knows what those circumstances are, her simplicity desire will not affect her preference.

Thus, seemingly, if I construe ‘conform to’ as ‘follow for a period’, Simplicity (which requires that rational moralists always have voiceable preferences for conformity to simpler norms in proportion to the strength of their simplicity desires) will not obtain. Nor, seemingly, will Invariance, for A will voiceably prefer that different moralists should follow norms of differing degrees of simplicity, depending on whether or not she believes that they will encounter the same sorts of circumstances throughout the relevant period.

There is another problem with proposal (2), which it shares—as far as I can see—with any proposal that allows A’s simplicity desire to significantly affect the content of her preferred norms. I call this the Uneven Compromise problem. But I will explain this problem in my discussion of proposal (3) rather than here, because the initial problem that I raised for (2) would make the Uneven Compromise problem harder to state clearly as an objection to (2).

(3) ‘Conforming to N’ as following N necessarily. According to proposal (3), a moralist A has a voice in what comprehensive norm-partition another moralist B necessarily follows, and her simplicity desire (vis-à-vis B) is a desire that the norms that B necessarily follows should be simpler. B does not satisfy A’s voiceable preference if he follows her best-preferred norms but would not have followed them had his circumstances been different.

Having a voice in what an agent necessarily does is apparently pointless. Presumably, if the agent necessarily follows norms N, then to demand that he should (necessarily) follow a different set of norms is to demand the impossible. So when we say that A has a voice in what norms B necessarily follows, the ‘necessarily’ must be understood in such a way that is possible for B to have ‘necessarily’ done otherwise. ‘Necessarily’ here, in both instances, must mean ‘necessarily, holding some set of token facts of a certain type (call them F-facts) fixed’, and ‘possible’ must mean ‘possible, without holding any token set of F-facts fixed’. This might suggest that A’s voice concerning which norms B should necessarily follow is really a voice concerning what the token F-facts should be. But this is not what I want to say. Rather, her demands do not specify a value for the F-facts; they are simply demands that he should necessarily follow certain sets of norms, and the fact that ‘necessarily’ here means ‘necessarily,
holding some set of token F-facts fixed’ is what explains why she is not, strictly speaking, demanding something impossible.

It must somehow be guaranteed that, if A is rational, the norms that A demands that B should follow are insensitive to F-facts: they must not give different prescriptions depending on the F-facts. The terms of the moral trust network could, I suggest, bring this about by stipulating that F-facts do not fall under admissible ground-types for members of the first order of the moral trust network. What are ‘F-facts’? I suggest that they are facts about the strengths of B’s trust-based desires, and about his actual or perceived trust-based obligations.

(As a general matter, it makes sense that moralists’ norms should be insensitive to entrusted parties’ trust-based motivations. It is by influencing these motivations that moral judgements have the potential to change moralists’ behaviour. By demanding that a moral entrusted party should φ, a moralist seeks to convince him that his trust-based obligations require him to φ, and so bring it about that his trust-based motivations move him to φ. Demanding that a moralist should follow norms prescribing φ-ing iff you have a trust-based motivation to φ would thus be ineffectual. Demanding that he follow norms requiring φ-ing iff you don’t have a trust-based motivation to φ would be self-undermining (and would have about it a whiff of the paradoxical). It would not serve the purpose of the moral trust network for its terms to allow that moral demands might be sensitive to entrusted parties’ judgements of trust-based obligation, or to their trust-based motivations.)

Unlike proposal (1), proposal (3) avoids making A’s simplicity desire too easy to satisfy. It is not always true that there is a simple comprehensive norm-partition that prescribes the actions that B would necessarily perform—that he would perform in every relevant possible circumstance. If A lights a cigarette in circumstances C but would have done a lot of very different things in a lot of circumstances that were very similar to C, then the comprehensive norm-partition that he necessarily follows (holding F-facts fixed) will probably not be as simple as it would have been (had he had a strong trust-based desire that moved him to follow a simple set of norms, say). Thus if A is rational, her simplicity desire will affect the content of her preferred norms.

Unfortunately this proposal seems to place excessive demands on moral entrusted parties. It makes violating A’s trust too easy for B. He can follow A’s preferred norms to the letter, and yet violate her trust because he would have defied those norms in some counterfactual circumstance (even with his moral trust-based desires and altruistic
desires still intact). Trust violations would be too common. It would not make sense for moralists to trust each other to be so robustly accommodating, nor for entrusted parties to commit to be.

To be sure, if A prefers that B should conform to comprehensive norm-partition N, B might avoid violations of A’s trust resulting from his counterfactually defying N, by forming discretion-licencing preferences that licence him to counterfactually defy N. B’s discretion-licencing preference, let’s say, would be a preference that B should follow a certain comprehensive norm-partition necessarily. If B necessarily follows this comprehensive norm-partition, then he doesn’t violate A’s trust by counterfactually transgressing N. As long as there is no point in time at which B counterfactually violates both his preferred norms and A’s, he will live up to A’s trust. But this still makes violating A’s trust too easy. Few of us are perfect, but far fewer are robustly perfect. B is bound at some point in time to be such that he lacks the trust-based motivation (specifically, the discretional desire) to counterfactually act on the most demanding prescriptions that his preferred norms and A’s preferred norms have in common, and then he will violate A’s trust, even if he has sufficient motivation to follow them in the actual world.

The problem is worse than it appears on the characterization I have just given it, because I have ignored two facts. First, A will not be the only moralist making demands of B; all of B’s partners in the moral trust network are sure to be making conflicting demands of him. So if B ever defies his own preferred norms qua moral entrusted party, he is sure to violate someone’s trust. Second, it ignores the fact that, if B is undecided about what to do in some circumstance C—which according to fiduciary non-cognitivism means having utilities that do not discriminate between comprehensive norm-partitions that give incompatible prescriptions in C—then he will violate someone’s trust regardless of the strength of his trust-based motivations, even if he is never in C. This is because C is a possible circumstance in which B cannot follow every other moralist’s preferred norms, and B lacks a voiceable preference that licences him to counterfactually resist following their norms in C. So proposal (3) has the result that B can violate trust merely by being undecided between norms. And being undecided between norms will be a common condition, because, as will become apparent later, the reasoning involved in forming voiceable moral preferences is difficult.

Proposal (3) also faces the Uneven Compromise problem that I alluded to above. This problem apparently arises for any proposal that predicts that simplicity desires as
well as altruistic desires will affect moralists’ preferences. It is the main reason why Simplicity and Invariance are hard to predict simultaneously. Suppose that moralist A is forming voiceable preferences as to which comprehensive norm-partitions moralists B1 and B2 should conform to, and that A’s preferences are being influenced by her desires that B1 and B2 should conform to simpler norms. The influence of this simplicity desire means that A will form preferences for norms conformity to which would not maximize the satisfaction of her altruistic desires. I’ll refer to the extent to which conformity to the prescriptions of a comprehensive norm-partition would satisfy a moralist’s altruistic desires as that norm-partition’s fit with her altruistic desires, or simply its fit. I will also call the extent to which conformity to a prescription of a norm-partition satisfies a moralist’s altruistic desires, that prescription’s fit with her altruistic desires. The influence of A’s simplicity desires compromises the fit of her preferred norms; it means that in forming her preferences she must make trade-offs between simplicity and fit.

There will be different ways for A to achieve simplicity, for these compromises could be made to different ‘parts’ of the comprehensive norm-partition. The fit with her altruistic desires of various different prescriptions could be sacrificed to achieve a simpler comprehensive norm-partition. And it might serve A’s weighted desires better for B1 and B2 to follow, at a time t, comprehensive norm-partitions that are compromised in different ways, and which are thus mutually incompatible. This is because A might know that B1 and B2 will occupy different circumstances at t, and hence will only have to perform actions that are relevant to those circumstances. So it will serve A’s preferences for B1 to follow a comprehensive norm-partition whose prescriptions in B1’s expected circumstances perfectly fit with A’s altruistic desires, and which give compromised prescriptions only in circumstances that B1 is unlikely to find himself in at t. Likewise, it will serve her preferences for B2 to follow a comprehensive norm-partition at t whose prescriptions in B2’s expected circumstances perfectly fit with A’s altruistic desires, and which trades off fit for simplicity elsewhere. So, immediately, before t, A will have voiceable preferences that B1 and B2 should follow incompatible comprehensive norm-partitions. But of course, if this is possible, then Invariance will not obtain. I call this the Uneven Compromise problem, for it arises because, when compromising the fit of her preferred comprehensive norm-partition to improve
simplicity, A has the option, so to speak, of concentrating these compromises in some parts of it rather than others.\textsuperscript{167}

The problem applies intertemporally as well as between agents. As B1’s expected circumstances change, A’s preferences concerning which norms he should follow will change accordingly. So if A knows that at $t_1$ B1 will be on board a sinking ship among children who will drown if he doesn’t help them on to a lifeboat, A might voiceably prefer, just prior to $t_1$, that he follow norms requiring him to always maximize expected life-preservation; and if A knows that at a later time $t_2$ B1 will have returned to his job as a journalist, and is in a position to expose corruption at the cost of his job, A might prefer, just before $t_2$, that he conform to norms requiring him to always maximize truth-telling. These might be the simplest comprehensive norm-partitions prescribing the actions, among those available to B1 at $t_1$ and $t_2$, that maximally satisfy A’s altruistic desires.

This sort of temporal fluctuation in A’s preferences is not, strictly speaking, incompatible with Invariance, but its possibility would undermine the comparative method of abuse detection in the same way that possible failures of Invariance would. To facilitate this method, a rational moralist’s preferences ought to be fairly stable over time, so that changes in a moralist’s demands (etc.) over time in similar cases will be evidence of ‘double standards’. We might call this desirable state of affairs \textit{Temporal Invariance}.

Thus far I have described the Uneven Compromise problem in a way that is supposed to be neutral between proposals (1)-(3), so it may not be immediately clear that the problem applies to proposal (3). But when we cash out talk of ‘conforming to norms’ as ‘following norms necessarily’ the problem remains. A will voiceably prefer that B1 necessarily follow a comprehensive norm-partition that is optimized for his actual situation, and that B2 necessarily follow a comprehensive norm-partition optimized for his.

This completes my survey of hypotheses concerning the kinds of decisions in which moralists have voices has such. All of these hypotheses have serious maladies. I will now consider which of them can be cured.

\textsuperscript{167} A dearth of vocabulary for describing the process of forming preference makes it is hard to avoid talking as though A were \textit{choosing} which norms to prefer, though the process is in fact non-voluntary.
Let’s start by addressing the Uneven Compromise problem. The problem is that, given all I’ve said so far, a moralist A will seemingly prefer that a moralist B should conform to different norms depending on his circumstances, because she will prefer that he conform to comprehensive norm-partitions that trade simplicity for fit only in their prescriptions concerning circumstances in which B is unlikely to find himself.

We can solve this problem by postulating that no proposition bearing on the contingent likelihood that B finds himself in one circumstance rather than another falls under an admissible ground-type for A; in other words, all such propositions are fundamentally inadmissible. This forces A to form her preferences concerning what norms B should conform to as if she were in complete ignorance about B’s contingent circumstances—as if she were behind a ‘veil of ignorance’, we might say.168 Grounds bearing on the contingent likelihood that she will find herself in one circumstance rather than another are also fundamentally inadmissible for A. So, assuming the strength of her simplicity desire is the same for her qua entrusted party as it is for her qua entrusting party, and assuming she is rational, she will voiceably prefer that she should conform to the same norms that she prefers (qua entrusting party) that others should conform to. Given these postulations, the Uneven Compromise problem shouldn’t threaten Invariance.

It is not available to the fiduciary non-cognitivist to hypothesize that the Uneven Compromise problem would be solved in exactly this way. As I mentioned above, moralists’ (‘effectively’ weighted) simplicity desires are derived. They depend on contingent beliefs. Some contingent beliefs must therefore be admissible behind the ‘veil of ignorance’ if simplicity desires are to affect moralists’ voiceable preferences. I discuss simplicity desires further, though inconclusively, in Section 6.6. But the fiduciary non-cognitivist can, I suspect, posit that the Uneven Compromise problem would be solved by a slightly adulterated version of the ‘veil of ignorance’ solution adumbrated above. (In what follows, I shall largely ignore complications to the veil-of-ignorance story arising from the derived character of simplicity desires.)

The Uneven Compromise problem was a general problem, facing more than one version of my hypothesis that moralists have voices concerning other moralists’ choices

of which norms to conform to. What of the problems specifically facing the particular versions? I am not aware of any way to solve the problem I pointed out for proposal (1), nor the problem proposal (3) gives rise to of making trust violations too easy. But I think that the maladies of (2) can be remedied to an extent that is sufficient to make (2) viable.

The distinctive problem facing (2) was that it did not predict that a moralist’s simplicity desire would influence her preference in a constant and reliable way. Suppose that moralist A has a voice in which comprehensive norm-partition moralist B should follow over a period \( d \), and A knows that B’s circumstances will remain the same throughout \( d \). A’s simplicity desire will be satisfied perfectly well as long as B responds to his circumstances in the same way throughout \( d \), for there is sure to be some very simple comprehensive norm-partition that he counts as having followed by so responding. So A’s simplicity desire will not dispose her to voiceably prefer that B should follow simpler norms over \( d \), as long as the norms she prefers that he should follow on the basis of her altruistic desires alone are norms that always require him to respond to his unchanging circumstances in the same way, as they surely will be. Simplicity will thus not obtain, and since the effects of A’s simplicity desire on her voiceable preferences in different relationships will vary depending on her beliefs about the likelihood that other moralists will find themselves in the same circumstances throughout \( d \), Invariance will not obtain either.

To solve this problem I don’t need to say much that is new. My hypothesized solution to the Uneven Compromise problem has the consequence that grounds bearing on the likelihood that moralists would find themselves in some circumstances rather than others are inadmissible. To solve the Uneven Compromise problem, the only grounds that really need to be fundamentally inadmissible for moralist A when A is forming a voiceable preference concerning what moral entrusted party B should do at a future time \( t_1 \) are those bearing on the probability that B will end up in some circumstances rather than others at \( t_1 \). However there is little further cost in postulating that grounds bearing on the probability that B will find himself in one circumstance rather than another from \( t_1 \) to a later time \((t_1+d)\), where ‘\((t_1+d)\)’ is the time that is later than \( t_1 \) by period \( d \), are also fundamentally inadmissible for A when A forms her preference about what B should do at \( t_1 \). And once we do this, we push facts about whether B will be in the same circumstance from \( t_1 \) to \((t_1+d)\) behind the veil of ignorance. A must always form her voiceable preferences as if she were in ignorance of
the likelihood that moralists would find themselves in the same circumstances throughout intervals of time that commence at the time of the decisions her preferences apply to, and that are no smaller than \( d \). This being so, her simplicity desire will presumably always affect her preferred norms to some degree, and always to the same degree, if she is rational.

Regardless of their contingent circumstances, moral entrusted parties will virtually always form their discretion-licencing preferences for options in the same sorts of decisions, behind the same sort of ‘veil of ignorance’. B will form preferences concerning which comprehensive norm-partitions to follow for the same interval \( d \), and grounds bearing on the likelihood that he will be in one circumstance rather than another over the course of \( d \) are fundamentally inadmissible for him. So Simplicity and Invariance will obtain.

Note that if we posit that precisification (2)’s problems would be solved by this means, we can posit a value for \( d \) that is as short as we like, as long as \( d \) is real-valued. A must form her preferences as if in ignorance of B’s contingent circumstances over \( d \) (or over the relevant \( d \)-long period, whenever its starting point might be), and if she were ignorant in this way, she would have no way of knowing how many different circumstances B would find himself in over \( d \), even if \( d \) were a tiny fraction of a second. A must thus form her preference as if, for all she knew, B might change circumstances radically from one moment to the next, like a spirit possessing now one body, now another, just long enough to make a decision (with lasting consequences) in each new circumstance.

Why would we want to postulate that \( d \) is brief? Here’s one reason: B might sometimes ‘change his mind’ qua moralist. Whether due to a change in his reasoning or a change in the strengths of his weighted desires, he might form discretion-licencing preferences to follow a comprehensive norm-partition that differed from the comprehensive norm-partition he preferred in the past. Suppose that at \( t_1 \) B prefers to follow norms \( N_1 \) over the period \((t_1 \text{ to } (t_1+d))\), that at \( t_2 \) B prefers to follow norms \( N_2 \) over \((t_2 \text{ to } (t_2+d))\), and that \( t_2 \) occurs before the time \((t_1+d)\) has elapsed. In the time between \( t_2 \) and \( t_1+d \) B is presumably obliged to follow both \( N_1 \) and \( N_2 \), which may be impossible. It would be unfortunate if B could easily violate trust in virtue of being obliged to follow incompatible norms as a result of ‘changing his mind’. One way to minimize this possibility is to make \( d \) too brief to allow these sorts of conflicting obligations to be at all consequential.
I hypothesize that the terms of the moral trust network realize Simplicity and Invariance in accordance with proposal (2) in conjunction with the remedy to (2)’s defects suggested above. These terms give moralists voices in one another’s choices of which comprehensive norm-partitions to follow over an exceedingly brief though real-valued period of time, which I will henceforth refer to as ‘d’. They give weight to moralists’ desires that moralists should follow simpler comprehensive norm-partitions over these periods. And they make grounds concerning moralists’ contingent circumstances over those periods inadmissible. Because the period over which moralists must follow their preferred norms is so brief, I will generally ignore it henceforth when describing the choices in which moralists as such have voices; I will merely say that moralists have voices in choices amongst which comprehensive norm-partitions to follow.

6.4 Preferences for Conformity to Non-comprehensive Norm-Partitions

6.4.1 Derived Voices and Norm-Indecision

I must postpone my explanation for the value of simplicity a bit longer. First, I need to examine the questions of whether and under what conditions a moralist may have a voiceable preference that anyone should follow a non-comprehensive norm-partition. My answers to these questions will have implications for the account I give of moral judgement and moral indecision. A new piece of terminology will be helpful in what follows: to avow norms or a norm-partition is to voice a preference for conformity to those norms or that norm-partition.

Let us now return to the questions I broached at the beginning of this section. In 6.2.1 I made some hypotheses concerning the supervenience of live options on live atomic sub-options. It is a consequence of these hypotheses that moralists have voices in moralists’ choices of which non-comprehensive norm-partitions to follow, among those that they presently can follow, because they have voices in moralists’ choices of which comprehensive norm-partitions to follow. To reiterate the relevant claims: the terms of entrusted concern relationships specify which decisions parties have voices in by specifying which atomic sub-options they have voices in; a party has a voice in any choice between two mutually incompatible options that are both realized by live atomic sub-options; and the weighted utility of non-live atomic sub-options is irrelevant to a determination of a party’s voiceable preferences among live options. For moralists, the
option of following a comprehensive norm-partition N is a live atomic sub-option of the option of following any non-comprehensive norm-partition that N entails. So a moralist has a voice—a derived voice—in any choice of a moralist between following one or the other of two incompatible non-comprehensive norm-partitions.\textsuperscript{169}

This means that, in a decision between performing actions \( \varphi \) and not-\( \varphi \) in a circumstance C, a moralist is entitled to voice a preference that another moralist should follow any norm-partition N that is entailed by her preferred comprehensive norm-partition, and which requires either \( \varphi \) or not-\( \varphi \) in C. This is because she has a derived voice in the decision to follow or not follow N, and the decision to \( \varphi \) or not-\( \varphi \) in C is, among other things, a decision to follow or not follow N. (If a preference for following a non-comprehensive norm-partition has no bearing on any decision that is under deliberation, it can still be voiced, but only as a preference in a counterfactual or conditional choice.\textsuperscript{170}) This is a major part of how I propose to explain the fact that moralists making public moral judgements don’t always do so by espousing comprehensive moral theories, but sometimes merely judge that some particular actions, or relatively small classes of actions with common properties (e.g. abortion), are right or wrong.

\textit{Obligations to accommodate} preferences for following non-comprehensive norm-partitions are explained by obligations to accommodate preferences for following comprehensive norm-partitions. (That is, the range of choices in which moralists have obligations to accommodate preferences includes choices among non-comprehensive norm-partitions, because it includes choices among comprehensive norm-partitions—I am talking about the scope rather than the source of moralists’ obligations here.) I have said that, in general, an entrusted party’s obligation to accommodate a preference voiced by an entrusting party in exercise of a derived entitlement is explained by his (the entrusted party’s) obligations to accommodate the voiceable preferences she has in his choices among live atomic sub-options—specifically, those of her voiceable preferences

\begin{footnotesize}
\textsuperscript{169} Bear in mind that to follow a non-comprehensive norm partition is to perform only actions that it requires. Thus whereas a moralist always has the option of following a comprehensive norm-partition, a moralist does not have the option of following a non-comprehensive norm-partition if there is no action available to her that falls within its extension.

\textsuperscript{170} Actually, it is probably better to regard all preferences concerning which non-comprehensive norms to follow in future choices as conditional, i.e. as preferences for follow norms N when in circumstances C. No one can know the agent’s future circumstances with perfect certainty, so voicing unconditional preferences for following non-comprehensive norm-partitions risks making demands of moralists that they cannot possibly live up to.
\end{footnotesize}
for the live atomic sub-options that explain the preference she voices in exercise of the
derived entitlement. Suppose that a moralist B can follow a non-comprehensive norm-
partition N—i.e. an action that N requires is available to him—and another moralist A
voices a (rational) preference that B should follow N. B is obliged qua A’s partner to
follow N, assuming he is, because he is obliged to accommodate A’s voiceable
preference in his choice of which comprehensive norm-partitions to follow, and the
option of following the comprehensive norm-partition that A best prefers is an atomic
sub-option of following N.

Similarly, a preference voiced by an entrusted party for a non-atomic option is
discretion-licensing, if it is, only derivatively, because B has a voiceable preference to
choose some live atomic sub-option that explains it. Thus if B voices a preference not to
follow N, and this preference is indeed voiceable, it is discretion-licensing because B
has a fundamental voice in his choice of which comprehensive norm-partition he should
follow, and, among all comprehensive norm-partitions, following the comprehensive
norm-partition that he best prefers to follow is a sub-option of following N.

The fact, noted above, that the weighted utility of non-live atomic sub-options is
irrelevant to the determination of a party’s voiceable preferences among live options has
the consequence that a moralist cannot necessarily scrupulously avow whichever non-
comprehensive norm-partition best satisfies her weighted desires. It implies that she
cannot voice a preference for conformity to a non-comprehensive norm-partition that is
not entailed by her best preferred comprehensive norm-partition (or norm-partitions—I
explain what I mean by this in a moment). Were her preferences among non-
comprehensive norm-partitions not constrained in this way, she could simultaneously
have voiceable preferences to follow multiple incompatible norm-partitions of varying
extensions.

One reason why it is important that a moralist should have voiceable preferences
for conformity to non-comprehensive norm-partitions is that this means that she will
probably have some preference she can voice even if she lacks a voiceable preference
for conformity to any single comprehensive norm-partition. More often than not, a
moralist will lack a voiceable preference for conformity to any single comprehensive
norm-partition. In what follows I will explain why this is so, after I have explained why
lacking a voiceable preference for conformity to any single comprehensive norm-
partition does not rule out having a voiceable preference for conformity to some non-
comprehensive norm-partition.
Voiceable preferences are based on expected weighted utilities. (The ‘expectations’ that are relevant to the determination of a moralist’s expected weighted utilities qua member of the first-order moral trust network must be based on the probability of propositions only insofar as those propositions are admissible. Thus they cannot be affected by moralists’ beliefs about the contingent probabilities that moralists will find themselves in some circumstances rather than others at the moment of the demanded action, and over the ensuing period \( d \).) Expected utilities—even perhaps rational ones—can plausibly be indeterminate, perhaps because credences can be indeterminate, and I shall assume that expected weighted utilities are no different.\(^{171}\) If a party to an entrusted concern relationship has indeterminate expected weighted utilities, these weighted utilities may fail to yield a ranking among two or more options that have the highest expected utility for her. This may result in her failing to have a voiceable preference for any single live atomic sub-option. But that doesn’t mean that she can’t have a voiceable preference among non-atomic options. She has a voiceable preference for any non-atomic sub-option if her best preferred live atomic sub-option is a sub-option of that non-atomic sub-option, even if it is indeterminate which of her live atomic sub-options this is.

Recall Stevenson’s football case. Suppose (implausibly) that John’s live atomic sub-options of the option (P) *John plays football* are all and only the following: (F) *John plays forward*, (D) *John plays defence*, and (G) *John plays goalie*. And suppose further that John voiceably prefers both F and D to G, but John lacks a voiceable preference between F and D because he has indeterminate expected weighted utilities. Finally, suppose that John’s best preferred sub-option of the option (∼P) *John does not play football*, namely, (C) *John joins the chess club*, has a higher expected weighted utility than G, but a lower expected weighted utility than each of F and D. Thus (F > C > G) and (D > C > G). Does the fact that John lacks a voiceable preference between F and D mean that he lacks a voiceable preference for P over ∼P? It does not. For John prefers P to ∼P if and only if the sub-options he best prefers out of all the live atomic sub-options of P and ∼P is a sub-option of P. And the sub-option he best prefers out of all his live atomic sub-options is, determinately, a sub-option of P, even if it is indeterminate

\(^{171}\) In the moral reasoning case, the credences whose indeterminacy might explain indeterminacy in a moralist’s expected weighted utilities would primarily be credences in *a priori* propositions about the implications of following certain norms. I give examples of such propositions in 4.2 below.
whether this sub-option is F or D. So John has discretion to resist his mother’s preference that \( \neg P \), though he lacks discretion to resist her preference between F and D.

The following is true for the same reason. Suppose that a moralist S has expected weighted utilities that fail to discriminate between her best-preferred options of following comprehensive norm-partitions. That is, there is some set of comprehensive norm-partitions \( (N_1, \ldots, N_n) \) such that \( (N_1, \ldots, N_n) \) has more than one member, and for all comprehensive norm-partitions \( N_x, N_y \) is in \( (N_1, \ldots, N_n) \) just in case there is no comprehensive norm-partition \( N_y \) such that conformity to \( N_y \) has determinately higher expected weighted utility for S than conformity to \( N_x \). The options of following the members of \( (N_1, \ldots, N_n) \) are S’s best preferred live atomic sub-options. Consequently, S will, for any norm-partition O such that following O necessarily realizes the option of following \( N_1 \), and \( \ldots \), and the option of following \( N_n \), have a conditional voiceable preference for the option of following O when possible. Which is to say that S will, for any norm-partition O such that O is entailed by each of \( N_1 \), and \( \ldots \), and \( N_n \), have a conditional voiceable preference for following O when possible.

Amongst these entailed non-comprehensive norm-partitions, I will call the non-comprehensive norm-partition with the largest extension S’s maximal decided norm-partition. Every moralist who has any voiceable preference at all has a maximal decided norm-partition. If she has a single, determinately best-preferred comprehensive norm-partition, this will count as her maximal decided norm-partition. Intuitively speaking, a moralist’s preference for conformity to her maximal decided norm-partition is the ‘strongest’, ‘most demanding’, and ‘most extensively discretion-licencing’ preference a moralist can scrupulously voice. On my usage, the norms entailed by a moralist’s maximal decided norm-partition are the norms that she ‘has’ and, collectively, are ‘her norms’. Any norm-partition that is compatible with a moralist’s maximal decided norm-partition but which is not entailed by it is a norm-partition about which she is norm-undecided; the option of following it is also one about which she is norm-undecided; the state of being undecided about a norm or option is one of norm-indecision.

6.4.2 The Difficulty of Moral Reasoning, and its Practical Implications

I have implied that moralists will find themselves in states of norm-indecision more often than not. This is partly because taking account of simplicity desires makes rational preference-formation more difficult. It means that a moralist cannot form her preferences among comprehensive norm-partitions in an entirely piecemeal way; she
must reason ‘holistically’. She cannot simply form her preferences by focusing on one area of conduct, then moving on to another, ignoring the results of her earlier ruminations, and ignoring areas of conduct that she has not considered yet. Rather, in forming preferences concerning any norm-partition, she must always be questioning whether or not there is a comprehensive norm-partition which entails it, and which is preferable to all rival comprehensive norm-partitions because it achieves a better overall trade-off between simplicity and fit.

The enterprise of forming rational preferences among norm-partitions is difficult both because it requires trade-offs between (‘holistic’) simplicity and fit, and because it requires an imaginative effort to grasp all of the implications of following a norm-partition whose extension is large. It may not immediately occur to someone who entertains a policy of always returning to people what they are owed that this would require returning an axe to its owner if he were insane and violent. It may not immediately occur to someone that a policy of maximizing happiness would in some circumstances require judicial murder. Thinking up sets of norms, forming *a priori* beliefs about their implications, and proportioning one’s expected weighted utilities for the options of following those norms to those *a priori* beliefs—this is what moral reasoning is all about according to fiduciary non-cognitivism. It is a difficult task, and one that many moralists would never complete to their satisfaction. Plausibly, it would not be uncommon for moralists to have expected weighted utilities that did not discriminate among their best-preferred comprehensive norm-partitions.

On my theory, moral reasoning is means-end reasoning of an apparently unusual sort. One undertakes it as if one were in (almost) perfect ignorance about the contingent probabilities that the world at the time of action would be one way rather than another. It might be doubted that there is any way to instrumentally reason well in such a state of ignorance, or that, if there were, humans would have any hope of doing it well, given their cognitive limitations. It requires a consideration of an uncountably infinite range of possibilities. In a way this is not unusual—there is an uncountably infinite number of ways the world might be in any real-world choice situation. But usually, it might be thought, we can get away with taking account of only a small and finite subset of these possibilities—those that meet a certain threshold of probability, one that is larger the more time we have to deliberate. This plausibly allows us to make tolerably good choices most of the time despite our brains’ limited processing power. But for moralists the contingent probabilities of different ways the world might be are treated as being
completely unknown. So if a moralist limits the range of possible circumstances that she considers to a finite and manageable set, she cannot do this on the basis of their contingent probability. We might worry that she must choose the membership of the set in way that is utterly arbitrarily.

I need to acknowledge that my account of moral reasoning raises problems of this sort, but I cannot attempt to solve them. This would be beyond my expertise, and would require too great a deviation from the main story I wish to tell. I do find it intuitively plausible that we can reason more or less well in situations of perfect uncertainty about the contingent facts. Suppose that you were forced to choose, in perfect ignorance of your future circumstances, and on the basis of purely self-regarding desires, some set of rules R such that (i) the complexity of R does not exceed a certain threshold, (ii) R can be stated without the use of proper nouns or indexicals, (iii) other agents would follow R in their interactions with you, and (iv) you would be forced to follow R in your dealings with others. It seems obvious to me that you could recognize that choosing some sets of rules, among those meeting criteria (i) and (ii), would be instrumentally better or worse for you than choosing others. Choosing rules requiring agents to maximally frustrate each other’s preferences would be worse than choosing rules requiring them to forbear from harming each other, for instance.

Let us return to the topic of norm-indecision. It is plausible that moralists would be norm-undecided about which norms to follow in a good many circumstances, but what are the practical implications of this uncertainty? A moralist as such can’t demand that others perform any specific actions in these circumstances, not without abusing other moralists’ commitments to live up to her trust. She might be able to demand that other moralists avoid performing actions in these situations though, for her maximal decided norm-partition might at least forbid certain actions in the circumstance, even if it did not require any actions. What would the moralist herself be obliged to do if she were in these circumstances? There is nothing that she could do that would not involve a trust violation, except to (very quickly) form a voiceable preference and act on it. In principle, she could choose an option φ if all other moralists either preferred that she should follow norms requiring φ, or had waived their preferences that she should follow norms forbidding φ, or were norm-undecided about whether to follow norms forbidding φ in her circumstances. This will never happen in practice however.

I will not present my account of moral uncertainty until Chapter 8. But, on the rough assumption that what I am calling norm-indecision is moral indecision, the
implications I have just noted at least seem to explain some elementary facts about our moral lives. If a moralist knows that she will in the future have to make a decision that she is presently morally undecided about, she will feel that she had better figure out what to do beforehand. If someone is unsure what morality requires of her in her present situation but must act regardless, this is not usually a circumstance that she will welcome.

6.4.3 The Salience of ‘Norm-Principles’

I’ll conclude this section by saying something about the reasons moralists would have for avowing various sorts of non-comprehensive norm-partitions. Two such reasons are obvious. First, a moralist may lack a voiceable preference for conformity to any specific comprehensive norm-partition (she may be in a state of norm-indecision); thus she could not scrupulously voice such a preference if she wanted to. Second, she may wish to be succinct and spare her audience a lot of information about her voiceable preferences that is probably not relevant to the decision under deliberation. This concern for relevance might lead us to expect that moralists would avoid avowing norms of any generality, and would instead just voice preferences for following those particular prescriptions of their norms that bore upon the present decision. But I think that there is a salient class of more general norm-partitions which moralists would often avow.

I call these norm-partitions ‘norm-principles’. I am discussing them not because they are vitally important to a satisfactory characterization of the moral trust network, but mainly in order to gesture at an explanation for the role of principles in moral thought and talk—having a norm-principle roughly corresponds to having a moral principle (a basic moral principle, not a derived one).\footnote{I shall not, in this dissertation, actually present an analysis of moral principles or judgements of moral principle, so what I say here is all I shall have to say on the matter. (I talked a great deal in Chapter 4 about ‘judgements of basic moral principle’, but this was just a way of signifying that I was talking specifically about judgements about what is necessarily right or wrong.)}

A norm-principle is a norm-partition whose extension consists of all and only those actions with a certain non-disjunctive set of (non-‘gruesome’, non-‘gerrymandered’) natural properties, and whose prescription function outputs only prohibitions, or only requirements. Norm-principles may be said to ‘subsume’ prescriptions made, and norm-partitions entailed, by (more general) norm-partitions that entail them. For all norm-principles L, all norm-partitions O, and all comprehensive
norm-partitions $N$, $L$ subsumes $O$ with respect to $N$ if and only if $L$ entails $O$ and $N$ entails $L$. I will be particularly interested in what I call ‘maximally general norm-principles’. For all norm-principles $L$, and all norm-partitions $N$, $L$ is a *maximally general norm-principle with respect to $N$* if and only if $N$ entails $L$, and there is no norm-principle $L^*$ such that (i) $N$ entails $L^*$, (ii) $L^*$ entails $L$, and (iii) $L$ does not entail $L^*$. Moralists can ‘have’ norm-principles. For all moralists $S$, and all norm-principles $L$, $L$ is one of $S$’s norm-principles if and only if $L$ is entailed by $S$’s maximally decided norm-partition. For all moralists $S$, and all norm-principles $L$, $L$ is one of $S$’s maximally general norm-principles if and only if, for all $N$ such that $N$ is $S$’s maximally decided set, $L$ is a maximally general norm-principle with respect to $N$.

Moralists, I contend, would often have an interest in avowing maximally general norm-principles, rather than avowing particular prescriptions on the one hand or comprehensive norm-partitions on the other. Broadly speaking, moralists’ simplicity desires will lead them to prefer comprehensive norm-partitions whose prescriptions are such that the smallest number of norm-principles will suffice to subsume them all. Since general norm-principles (i.e. those with large extensions) can subsume more prescriptions, the fact that a comprehensive norm-partition entails a general norm-principle (the more general the better) tends to speak in its favour. A moralist who wants to demand conformity to, or claim discretion to follow, a particular prescription in a case under deliberation, will often want to establish that the preference she voices is rational given her simplicity desire. This being so, she has an interest in communicating to other moralists the fact that she has a norm-principle (the more general the better) which gives that prescription. So we might expect her to voice a preference for following a maximally general norm-principle rather than merely for following the particular prescription that is most immediately relevant to the case.

This does not involve communicating a large amount of information of doubtful relevance to her audience. For one thing, it is economical to avow a norm-principle; it involves communicating something simple: a set of natural properties and a prescription applying to all actions with those properties. For another thing, her audience will often be interested in knowing whether her voiced preference is rational, because if it isn’t, any moralists whose preferences oppose hers may be able to change her mind and/or discredit her by pointing out flaws in her reasoning. The generality of her norm-principle (as well as its fit with her altruistic desires) bears upon its rationality, given that she has an effectively weighted simplicity desire.
Again, by arguing that moralists would often avow what I am calling ‘norm-principles’, I hope to account for the evident importance of principles in moral talk. One specific phenomenon that I must explain is that, when a moralist supplies grounds for a particular moral claim she has made, she will often cite a principle that makes that prescription. Or else she will cite contingent facts that explain why a principle (perhaps unstated) makes that prescription in the present case—which is also a way of implicitly citing a principle. This contrasts with the practices of parties to entrusted concern relationships such as I discussed in Chapter 1, who will cite as grounds facts that explain why an option would serve their weighted desires, rather than why it is required by a principle. I think that parties to these sorts of entrusted concern relationships on the one hand, and moralists on the other, are all trying to do the same thing, however. They are trying to show that their demands or claims to discretion are supported by their weighted desires, the ultimate source of their trust-based entitlements. The difference in the moral cases is that moralists’ weighted desires include simplicity desires, and citing the general norm-principles that lie behind one’s demands and claims is a way of indicating that one’s voiced preferences are duly influenced by one’s simplicity desires.

6.5 Advantages of Simplicity

6.5.1 How ‘Simplicity’ Aids the Comparative Method of Abuse Detection

I have explained how the terms of the moral trust network might bring about both Simplicity and Invariance, but I still need to explain the advantages of their doing so. To this end, I shall look at a case in which realizing Simplicity would plausibly make a difference to the likelihood that an entrusting party could detect abuse through a comparison of voiced preferences. I will consider how easily abuses could be detected in this case in the presence and in the absence of a justified expectation that voiceable preferences will be influenced by simplicity desires. To make the contrast cleaner, I shall assume throughout that moralists have voices in one another’s choices of which comprehensive norm_partitions to follow, rather than simply in their choices among actions, even though this would not really make sense if moralists did not have weighted simplicity desires. My explanation for the value of Simplicity relies mainly on facts about the amount of information that moralists communicate about their norms when they voice preferences—about the extent to which, by voicing a preference for
following a non-comprehensive norm-partition of a given extension, a moralist narrows down the range of more comprehensive norm-partitions she is likely to prefer.

Suppose first that Simplicity does not obtain; a moralist’s norms are thus as complicated to describe as her preferences based on her altruistic desires alone are. It is now $t_2$, and moralist Bob has just voiced a preference to follow norms requiring $\phi$ in his present circumstance $C$, in response to another moralist’s demand that he follow norms requiring not-$\phi$. It will serve Bob’s private (i.e. non-altruistic) desires to choose $\phi$. But the two options are such that a person might well prefer one or the other on the basis of her altruistic desires alone, depending on the relative strengths of those desires, so Bob’s preference may or may not have been biased by self-interest.

A fellow moralist, Sue, wants to determine now whether Bob’s voiced preference is his genuine voiceable preference, or whether he is trying to get away with defying his discretion-licencing preferences. Sue has a fair degree of familiarity with Bob and the preferences he has voiced qua moralist. Nobody had any idea this morning, at $t_1$, that Bob would end up in $C$. So if Sue had asked Bob this morning whether his norms required moralists to $\phi$ or not-$\phi$ in $C$, his answer would have been unbiased, and Sue would now be in a position to determine, with a high degree of confidence, whether the choice Bob made at $t_2$ was biased. And perhaps Sue did. More generally, perhaps Sue gathered enough information prior to $t_2$, so that she might now determine what she wants to know. Our question is, how likely is this?

The answer is, I think, that the information Sue needs is too specific for there to be much likelihood that she would have acquired it ahead of time. Sue might have asked Bob at $t_1$ or earlier whether his norms requiring $\phi$ or not-$\phi$ in $C$, but why would she have? Sue might know of a precedent of Bob’s choice situation, whose ‘altruistically relevant’ features (i.e. those features that might make a difference to the extent to which different options would satisfy altruistic desires) were identical to those of Bob’s choice situation, in which Bob voiced a preference qua moralist. But this is surely highly unlikely. Sue might know of a near precedent, which differed slightly in its altruistically relevant features, in which Bob voiced a preference for norms forbidding the $\phi$-like option. This is less improbable. But this knowledge by itself would not allow Sue to infer that Bob had been unscrupulous. Perhaps the slight differences in the case’s

173 Although of course, if Bob had earlier voiced a preference for following norms forbidding $\phi$, he probably wouldn’t be trying to get away with $\phi$-ing now.
altruistically relevant features was enough to explain the contrasting prescriptions of Bob’s norms.

The only circumstances under which Sue would reliably be able to gauge Bob’s scrupulousness is if she had at hand a fully worked-out and well-supported theory about the strength and content of his altruistic desires. And of course she could have such a theory at hand. She could have acquired the data for such a theory by taking note of the preferences Bob voiced, and also by asking Bob how he would prefer that entrusted parties act in hypothetical cases, and on what grounds. Or someone else might have done this work, and communicated the theory to Sue. But such research and theorizing would require a big investment of time and effort, and it is doubtful that this would have seemed worthwhile to anyone. So Sue will probably be none the wiser as to whether Bob’s voiced preference was biased. The problem is not that Sue could not have acquired the information she now needs, but that the information she now needs is too specific for there to be much likelihood that she would have acquired it, whether by accident or design.

It seems as though this problem would be partially overcome if Simplicity obtained. For then Sue would be justified in assuming that Bob’s norms (and those of moralists in general) were such that they could be subsumed under norm-principles that were relatively few in number and relatively general. The likelihood that Sue would be in a position to determine whether Bob’s voiced preference was biased would be greatly increased, because the facts she would need to have learned beforehand in order to do so would be far fewer in number, and of greater general interest.

Suppose that Bob’s demand is biased; his norms really forbid \( \varphi \). Because his norms are influenced by his simplicity desire, their prescriptions can be subsumed under a relatively small number of relatively general norm-principles. This is presumably true of the prohibition of \( \varphi \). The norm-principles that capture the prescriptions of Bob’s norms most economically include a somewhat general norm-principle forbidding all actions with some property \( \text{F} \), where \( \varphi \) has \( \text{F} \). To learn that Bob’s voiced preference to follow norms requiring \( \varphi \) is biased, Sue would need only to have asked Bob before \( t_1 \) what norm he preferred that people conform to, with regard to any \( \text{F} \)-action, and for Bob to have explained his preference regarding that action by citing the norm-principle forbidding \( \text{F} \). This is far more likely than that Sue would have asked about Bob’s preferences concerning the choice between \( \varphi \) and not-\( \varphi \) in \( C \) specifically, or some very similar choice, in which his private interests were not at stake. It is also not too
improbable that questions about F-actions might have arisen by accident—someone might have wanted to perform an F-action, and Bob might have opposed this, voicing a preference for following norms forbidding F. This is far more likely than that Bob’s choice happened to have had a precedent, where the earlier choice had the following properties: it was nearly exactly similar in its altruistically relevant properties to Bob’s choice at $t_2$, it was a choice in which Bob’s private interests were not at stake, and it was a choice concerning which Bob voiced a preference.

One reason why Sue is more likely to know Bob’s norms concerning $\phi$ before $t_1$ is that whenever he voices a preference qua moralist, other moralists enjoy discursive entitlements to demand that he say more about the content of his norms as a whole. They have these entitlements, either as members of the first- or second-order moral trust network, because parties to entrusted concern relationships in general are entitled to demand that their partners explain why they have the preferences they have voiced, to establish whether those preferences are genuinely voiceable, and to scrutinize the reasoning behind them. If Bob’s preferences are influenced by his simplicity desire, then the explanation for his voiceable preference concerning any particular action will be a holistic one, to which all the prescriptions of his maximal decided norm-partition are potentially relevant. Thus, if Simplicity obtains, these familiar discursive entitlements allow moralists to demand that Bob say more about his norms as a whole. This increases the likelihood that Sue would have learned of Bob’s norms concerning $\phi$ if she had previously heard him voice a preference in an unrelated decision.

In general, the obtaining of Simplicity would make it easier to narrow down Bob’s probable norms from the history of preferences he has voiced when his interests were not at stake, even if he never, before $t_2$, voiced a preference for following norm-principles forbidding actions with any of $\phi$’s properties. This is because, if Simplicity obtains, the number of different comprehensive norm-partitions a rational moralist could plausibly prefer is vastly diminished (though the number might still be very large indeed). Concerns of simplicity would induce moralists to iron out quirks in their norms that would otherwise have been points of difference between their norms and those of others. Thus Sue could more successfully glean skerrick evidence bearing on Bob’s scrupulousness if Simplicity obtained than if it did not.
6.5.2 Other Advantages

Simplicity, I think, has advantages beyond making opportunities for abuse-revealing preference comparisons more common. It also tends to make the norms that moralists avow ‘fairer’, in certain respects, when they are avowed insincerely. It has the result that when moralists abuse trust by voicing preferences that are biased to favour their private interests, they cannot so easily tailor the bias (so to speak) so that it benefits or harms only some people and not others. This means that when abuses happen, they tend not to be as disproportionately beneficial to the abusive and costly to the scrupulous as they would have been did Simplicity not obtain. This tends to make unscrupulousness less attractive and scrupulousness more attractive.

An unscrupulous moralist will typically want to appear to be acting in accordance with her norm-principles, so as not to make others hostile to her, even while performing actions that her true norm-principles do not permit. She will want to make public moral judgements that permit her to do things that serve her partial desires, by which I mean her desires specifically for the interests of her favoured parties (e.g. herself, her friends, her family) to be promoted, and for the interests of her disfavoured parties (e.g. her foes, her rivals, those she envies) to be thwarted. But she will also tend to want to publicly judge that altruistic-desire-frustrating actions are wrong and altruistic-desire-satisfying ones are permissible or required. This is both to preserve the appearance that her avowed norms are those she prefers qua moralist, but also to permit her to perform, without apparently exceeding her discretion, actions that better satisfy her altruistic desires: unscrupulous moralists have altruistic desires just as scrupulous ones do, so they would rather follow (and demand that others follow) altruistic-desire-satisfying norms, all things being equal.

And she won’t want her moral proclamations to be too obviously adulterated to permit her to be partial toward herself and her favourites. Suppose that the altruistic desires of an autocratic ruler, King Boniface, urge him to always punish murder with death, while his partial desires urge him not to execute his favourite, Sir Valiant, for committing murder. To create the appearance that his true norm-principles require sparing Valiant, it wouldn’t do for Boniface to avow norms requiring him to spare murderers just in case they are named ‘Valiant’. This would make his bias too obvious. Being named ‘Valiant’ is not an altruistically relevant property; there is thus no set of altruistic desires that Boniface might have on the basis of which he would rationally prefer that murderers named ‘Valiant’ should be spared, and that murderers with other
names should be executed. Rather Boniface will want to identify some highly
distinctive altruistically relevant feature of Sir Valiant’s case that could, for some
possible set of altruistic desires, explain why Boniface would judge it permissible to
spare Valiant from execution. For instance, suppose that Valiant is very courageous: he
once led a cavalry charge against a force three times as large as his. Boniface could
publicly judge it permissible to spare murderers of ‘exceeding valour’, specified to be
Sir Valiant’s level of valour and nothing less. This would allow him to spare Valiant
(and so satisfy his partial desires) but few other murderers, if any (and so satisfy his
altruistic desires), while acting in accordance with his moral claims.

Boniface could use the same strategy to spare all of his favourites from
punishment. Suppose that Lady Charity and Lord Fidelio are also favourites of
Boniface, and also took part in the murder Valiant committed. Boniface might avow
norms according to which Charity’s extraordinary charity and Fidelio’s extraordinary
loyalty pardon their involvement. He could do this sort of thing quite a bit without
making the idea that his avowed norm-principles were those he genuinely preferred on
the basis of his weighted desires too incredible, if those desires did not include a
simplicity desire. But if it could justifiably be presumed that he had a weighted
simplicity desire, the credibility of this idea would be less in proportion to the presumed
strength of that desire. The quirky exceptions to Boniface’s general norm-principle that
murder should be punished with death are just the sorts of things we should expect a
simplicity desire to militate against. So if Boniface is presumed to have a weighted
simplicity desire, his aim of avowing norm-principles that might plausibly be his
genuine norm-principles will incline him to either abandon all exceptions to the general
norm-principle, or include a single broad exception that will require him to spare all
three murderers. If he chose the latter course, the property that made murder pardonable
would probably be less distinctive than the properties of extraordinary valour, charity,
and loyalty that Valiant, Charity, and Fidelio exhibit, for it would have to be ones
shared by all three murders (or murderers). So Boniface’s norm-principle concerning
sparing murderers will probably apply to many people, not just his favourites.

From the point of view of members of the moral trust network, this is, on balance,
a ‘good thing’. Simplicity forces those who would avow norm-principles designed to
promote the satisfaction of their partial desires, to more heavily sacrifice the tendency of
these norm-principles to promote the satisfaction of their altruistic desires, and this
makes being unscrupulous less attractive. If Boniface wants to claim that it is right to
spare Valiant, Charity, and Fidelio from execution, he will probably have to claim that it is right to spare a lot of murderers whom he would rather were executed.

There is admittedly a drawback here: if Boniface’s altruistic desires are not eccentric, then this course, if chosen, will probably lead to a greater net frustration of most moralists’ altruistic desires, not merely his. And of course it is the purpose of the moral trust network to prevent such frustrations. But here are three reasons to think that the effect of Simplicity I have been describing is advantageous for moralists despite this problem.

First, Boniface will be inclined, when choosing which property of murderers spares them from execution, to pick some property such that sparing murderers with that property does not frustrate altruistic desires too much. Both his own altruistic desires and his desire that his avowed norms be credibly genuine require this. So the increase in altruistic-desire-frustration caused by knavish moralists like Boniface when Simplicity obtains shouldn’t be too great.

Second, there is some plausibility in the idea that the moral trust network will not persist unless moralists are able to trust one another to reciprocate their sacrifices. So it may be worth diminishing moralists’ incentives to shirk reciprocation, even if this means that altruistic-desire-frustration will be worse when they do. A compromised moral trust network may be better than none at all.

Third, when moralists insincerely avow norms as Boniface does, there is at least an upside, of a self-regarding sort, for other moralists who are affected by their actions. Moralists, and those to whom they are partial, will sometimes enjoy the favourable treatment that unscrupulous moralists would, absent Simplicity, only have granted to themselves and those to whom they were partial. Suppose that a moralist S among Boniface’s subjects, or a friend of S’s, commits murder. It would serve S’s partial desires for her or her friend to be spared from execution. If Simplicity obtains, there is a chance that she or her friend will enjoy the same indulgence that Boniface extends to his favourites, because she or her friend may happen to possess the not-too-unusual property that excuses murder according to Boniface’s moral proclamations. Thus Simplicity means that benefits that would only have been enjoyed by the unscrupulous and their favourites would be spread to many others, whether or not they were unscrupulous or had unscrupulous friends. This diffusion of benefits would lower the relative cost of being scrupulous.
6.6 The Difficulty of Explaining Simplicity Desires

Thus far I have explained why the terms of the moral trust network would be so contrived as to bring it about that moralists prefer conformity to simpler norms—which roughly means explaining why such a contrivance would serve moralists’ interests. I have also given part of an explanation of how this contrivance would work. First, moralists have desires that they and others should conform to simpler comprehensive norm-partitions, and the terms of the moral trust network count these desires as (effectively) weighted. Second, to ensure that the result of counting these desires as weighted is that moralists actually voice preferences for following simpler norms, and to ensure that their preferred norms do not fluctuate wildly as their beliefs about future circumstances change (the Uneven Compromise problem), the terms of the trust network give moralists voices only in one another’s choices concerning which norms to follow over a certain interval of time $d$, and make potentially Invariance-undermining facts about future circumstances fundamentally inadmissible for moral reasoners.

But my explanation for the importance of simplicity in moral theorizing will not be complete until I explain why moralists have simplicity desires in the first place. As of writing, I have failed to do so. In this section, I make some brief comments on why providing such an explanation is difficult, and on some of the costly implications an adequate explanation would likely have for fiduciary non-cognitivism.

The simplicity desire cannot be a basic desire that people should follow simpler norms. For my theory to be adequate, the desire must be a desire that humans almost invariably have, and I take it that basic desires only exist with this sort of reliability if they are the product of evolution. It seems that no satisfactory evolutionary explanation can apparently be given for the existence of a basic simplicity desire. For such an explanation to be plausible, simplicity desires would presumably have to have conferred adaptive advantages on their possessors. But while this would plausibly have been true at some stages in the history of our species, it would probably not have been true at the time required by an explanation for the desires’ existence.

The adaptive advantages would presumably be these: having strong simplicity desires makes one easier to trust qua moralist, and hence less likely to come under suspicion of abuse, and hence less likely to suffer the consequences of being thought to have violated trust (poor reputation, ostracism, punishment, and so on). But for these advantages to explain humans’ possession of simplicity desires, the moral trust network
would have to have been ‘up and running’ without the aid of any basic simplicity desire for many generations. Yet the terms of the moral trust network are the way they are, and so have features that would plausibly confer a fitness advantage on moralists with strong simplicity desires, only because they are designed to exploit moralists’ simplicity desires, as a means, chiefly, of reducing the epistemic obstacles to trust among moralists. So even if humans have evolved to possess basic simplicity desires (a view I would rather not be committed to), the existence of the moral trust relationship cannot plausibly depend on their possessing basic simplicity desires. Moralists must have, or must once have had, simplicity desires that are not basic, and so not explained by any special evolutionary story.

Derived desires, of course, depend for their rationality on beliefs—frequently on contingent beliefs. If derived desires based on contingent beliefs are to influence moralists’ voiceable preferences, then the facts that these beliefs are about must fall under admissible ground-types. So if the simplicity desires I posit are derived desires, as it seems they must be, some contingent facts will have to be admissible—there must be chinks and perforations in the veil of ignorance. This may or may not be a problem depending on the facts in question, for admissible ground-types can often be cunningly engineered to block inferences from contingent propositions that might threaten Invariance. But it is certainly a complication.

Any story I tell to explain moralists’ simplicity desires must meet the following three desiderata. First, simplicity desires must be desires for following simpler norms; they can’t merely be desires for voicing preferences or having voiceable preferences for following simpler norms. This is a perhaps obvious, but it is worth stressing. It is very easy to think of reasons why a moral entrusted party would want to voice preferences to follow simpler norms: he wants to be trusted by other moralists; he doesn’t want moralists to think that his avowed norms have been skewed so as to seemingly licence extra-discretionary actions; he wants his norms to be well-known so that when he acts on them in cases in which doing so serves his partial desires, he will not come under suspicion. It is tempting to think that these motivations would result in his having preferences for conformity to simpler norms, but they would not—at least not straightforwardly.

Second, the explanation must predict that a moralist will have the same norms qua entrusted and entrusting party, if she is rational and scrupulous. Here is a hypothesis that I have explored, which seemingly lacks the resources to meet this desideratum. A moral
entrusted party wants to *avow* simpler norms for the reasons given above. But he also wants to *follow* simple norms so that he can in the future avow simple norms without appearing to have abused trust by violating them in the past. It is these desires to follow simple norms so that he can avow them with an appearance of sincerity that are his simplicity desires. This story is odd and contrived, but perhaps it can be got to work. However, the described desire seems peculiar to entrusted parties. There is no reason to think that our moralist will have a desire qua entrusting party that other moralists should follow norms that are as simple as the norms that he desires to follow qua entrusted party, given the provenance of the latter desire.

Finally, an account of moralists’ simplicity desires should predict that the strength of these desires will not vary too much over time. For such variation would lead to variation in moralists’ norms, and if these are too mutable, moralists’ capacities to detect abuses via the comparative method will be undermined.

### Conclusion

The apparent rule-like character of moral requirements is, I have argued, explained by the following facts. Abuses of discretion, concealed by insincere voicings of preferences can sometimes be discovered through a comparison of voiced preferences past and present. Such comparisons provide better evidence of insincerity, where it exists, if variation among people’s voiceable preferences is inhibited somehow. One way in which the terms of the moral trust network can inhibit variation in moralists’ voiceable preferences is by (i) entitling them to voices only in moralists’ choices of which comprehensive norm-partitions to conform to, (ii) counting their desires that moralists should follow simpler norms among their weighted desires, assuming that they have such desires, and (iii) counting as fundamentally inadmissible any propositions a consideration of which might lead to their having voiceable preferences for moralists to conform to different comprehensive norm-partitions depending on their contingent circumstances. If the terms of the moral trust network have these features, moralists will tend to have voiceable preferences for conformity to simple principles that vary little from decision to decision. Thus I can explain the apparent rule-like character of moral requirements as resulting from the need to detect abuses of trust in a network of strangers. Or rather, I can do so if I can explain why moralists have would have ‘simplicity desires’, something I have so far failed to do.
I’ll now take a short break from theory-building to trumpet one of the main advantages of my theory. The theory plausibly has weaker commitments in the philosophy of mind than most non-cognitivist moral psychologies, because it can explain the changes in motivation produced by moral reasoning without relying on the thesis that people’s higher-order non-cognitive attitudes (i) can effect changes in their lower-order non-cognitive attitudes, and (ii) reliably do so in the right circumstances. Most contemporary non-cognitivists seem to be wedded to this thesis—to the view that our higher-order non-cognitive attitudes, as I shall put it, have _heft_ in determining our lower-order attitudes; they are committed to what I will call the _Heft_ thesis. By ‘higher-order non-cognitive attitudes’ (or ‘higher-order attitudes’ for short) I mean non-cognitive attitudes toward having other (‘lower-order’) non-cognitive attitudes—desires to have desires, for instance. Contemporary non-cognitivist theories have not typically entailed Heft, but they have relied on it for their plausibility. This reliance is on Heft unfortunate, and gives us reason to prefer fiduciary non-cognitivism to its non-cognitivist rivals.

As I have suggested, the main reason for non-cognitivists’ reliance on Heft seems to be that it gives them a way to explain the changes in moral motivation that sometimes result from substantive moral reasoning. By ‘substantive moral reasoning’ I have in mind the sort of moral reasoning that is associated with the term ‘reflective equilibrium’, that is not merely deductive, and that—to use familiar but metaethically loaded language—chiefly involves devising moral principles and testing them against intuitions about cases.

My main concern in this chapter will be to point out some of the disadvantages of a commitment to Heft, so that my theory’s advantage on this point will be more evident. I will first discuss some ways in which non-cognitivists have relied on Heft, giving
special attention to their accounts of substantive moral reasoning. Then I will examine the reasons for this reliance in the case of substantive moral reasoning. Thirdly, I point out some of the costs that come with commitment to Heft. Finally, I will spell out how fiduciary non-cognitivism eschews this commitment.

7.1 Examples of Expressivists’ Commitment to Heft

Let us consider some of the ways in which proposed expressivist theories have made use of higher-order attitudes, and why this makes their plausibility dependent on Heft. Simon Blackburn’s early work on expressivist semantics relied on the thesis to explain consistency and inconsistency among moral statements.174

Later theorists have also proposed accounts of the non-cognitive attitudes involved in moral judgement as higher-order attitudes. Allan Gibbard distinguishes moral from non-moral normative judgements by positing that the former are non-cognitive attitudes of ‘accepting’ norms prescribing feelings of guilt and resentment.175 Mark Schroeder suggests that expressivists should posit that wrongness judgements are attitudes of being for blaming for actions. The complexity of the proposed attitude-type is meant to provide a means to accommodate the various ways of negating a wrongness judgement report.176 It is not clear whether Schroeder is thinking of blame as an attitude or an action.177 If the former interpretation is correct, Schroeder is proposing that moral judgements be construed as attitudes toward other attitudes.

174 For instance, when characterizing mental states expressed by disjunctive sentences, Blackburn posited a non-cognitive mental state of being tied to a tree between two other mental states. This state has the following property: necessarily, someone who is tied to a tree between the state expressed by sentence P and that expressed by sentence Q and comes to have the state expressed by ~P is ‘committed’ to having the state expressed by Q. The states between which one may be tied to a tree include non-cognitive attitudes. (Blackburn, Essays in Quasi-Realism, 192-3; Schroeder, Noncognitivism in Ethics, 140.) Blackburn is thus committed to the view that the state of being tied to a tree is a hefty attitude, assuming that he thinks that people regularly succeed in making ‘(P or Q), ~P, therefore Q’-type inferences in cases in which P and Q express non-cognitive attitudes.
175 Gibbard, Wise Choices, 52.
176 Schroeder, Being For, 58-9.
177 For the former reading see Neil Sinclair, ‘Moral expressivism and sentential negation’, Philosophical Studies 152, no. 3 (2011): 409; for the latter, Schwartz and Hom, ‘Why the Negation Problem,’ 839-40. As Schwartz and Hom point out, Schroeder’s claim that his proposed moral attitudes are motivating suggests the ‘action’ interpretation. The attitude interpretation is supported by Schroeder’s claim (Being For, 58) to be borrowing from Gibbard’s account of moral judgement in Wise Choices, Apt Feelings, where Gibbard clearly has in mind attitudes toward norms prescribing attitudes (guilt, anger, etc.); and also by Schroeder’s entertainment of the idea ‘that it is impossible to both blame for murdering and blame for not murdering’ (Being For, 73), which would make more sense on an ‘attitude’
These proposals—assuming the ‘attitude’ interpretation of Schroeder—would have odd implications if they were not conjoined with Heft. Blackburn would be unable to explain how people actually make successful deductive moral inferences. Gibbard and Schroeder would be left with a view of moral thought and talk as a practically irrelevant activity, on which people reasoned and argued about when to feel guilt and anger, or when to blame; and then felt guilt and anger, or blamed, in exactly the way they always had regardless of the outcomes of their thinking and arguing.

Expressivists’ accounts of substantive moral reasoning tend also to depend on hefty higher-order attitudes. Expressivists have not given as much attention to substantive moral reasoning as they have to deductive reasoning, perhaps because expressivism is a semantic (or metasemantic) thesis, and the semantic claims of expressivists depend only on their accounts of deductive reasoning. But when they do discuss substantive moral reasoning, expressivists tend to propose that it involves bringing lower-order moral attitudes into a better fit with higher-order attitudes of some sort. These higher-order attitudes are typically said to constitute normative judgements of a higher-order sort. On some views the normative judgements are (higher-order) moral judgements; on others they are, or are at least analogous to, normative epistemic judgements.

Blackburn holds that some sets of moral attitudes are better than others, and that the value of moral attitudes is determined by moral criteria, judged from within a moral perspective. That is, when I wonder how I might improve, I have to think about deploying my current attitudes—there is no standing aside and apart from my present sensibility. … [W]hatever flaws I suspect are judged as flaws in the light of other concerns (Neurath’s boat). … I might certainly suspect flaws other than sheer inconsistency: immaturity, lack of imagination, bias, coarseness, and so on. Blackburthinks that the higher-order moral judgements involved in evaluations of our moral attitudes are not immediately judgements about those attitudes, but rather are fundamentally judgements about ourselves—about how we may be made better or worse. A few passages from Ruling Passions suggest that Blackburn endorses Heft.

---

reading. However we interpret Schroeder, his being for blaming for formulation was only proposed to illustrate a general strategy for solving the negation problem, not as a considered and specific view of how expressivists ought to characterize moral judgements.

178 Blackburn, Ruling Passions, 313; cf. 318.
179 See also Blackburn’s suggestion that we can ‘cleanse ourselves’ of ‘unfair’ non-cognitive attitudes (Ruling Passions, 309-310), his answer to his ‘Question 8’ (313-314), and what might
the most evocative of which is his approving characterization of a view of normative deliberation that he attributes to David Hume and Adam Smith:

We can start by presenting the rival theories of deliberation in terms of the Platonic model of a person as a ship. For Hume or Smith, the ship is worked by a crew, each representing a passion or inclination or sentiment, and where the ship goes is determined by the resolution of conflicting pressures among the crew. After one voice has prevailed, various things may happen to the losers: they may be thrown overboard and lost altogether, or more likely they may remain silenced just for the occasion, or they may remain sullen and mutinous, or they may continue to have at least some effect on the ship’s course. Some crew members may be more stable and durable, long-serving and forceful than others. Some, such as those labelled “prudence” and “industry”, may largely have a second-order role, to encourage or silence other, first-order crew members, such as “sloth” or “gluttony”.

This passage describes a number of ways in which desires can affect other desires. Among these, the possibility that desires might be ‘thrown overboard’ by other desires seems to be the sort of influence posited by Heft.

In *Wise Choices, Apt Feelings*, Gibbard writes, ‘Some norms govern the acceptance of norms. These include norms of coherence and they include norms that go beyond sheer coherence.’ Gibbard calls ‘norms that govern the acceptance of other norms higher order norms’, and later, ‘epistemic norms’. The normative judgements one regards as rationally required (in a broad sense of rationality) under certain circumstances will be those which the higher order norms one accepts prescribes in those circumstances. Gibbard thinks that the theory of reflective equilibrium may be thought of as a system of higher order norms. He doesn’t think that the only way in which we come to form or change our normative judgements is by forming lower-order norm-acceptance attitudes that ‘satisfy’, as it were, our higher order norm-acceptance attitudes (i.e. that are prescribed by the higher-order norms one accepts). But those of a moralist’s moral judgements that she regards as rational will be those that are prescribed by her higher order norms.

perhaps be construed as a suggestion that second-order desires ‘drive’ our ‘selection’ of first-order desires (255).

---

180 Blackburn presents the Humean and Smithian view of deliberation as a foil to the Kantian view that he opposes. *Ruling Passions*, 245ff.
According to James Lenman, ‘Moral inquiry … is, fundamentally a practical activity addressed to practical problems.’ The basic practical problem is ‘that of bringing the desires, aspirations and commitments we bring with us to the enterprise into desirable forms of coherence’, understood both intrapersonally and interpersonally.\textsuperscript{187} Such attitudes ‘are liable to conflict’ with each other. Intrapersonally, [first-order desires … can hardly do otherwise than conflict pervasively. But the project of reaching and reflectively endorsing an adequately coherent set of higher-order reflective commitments about which of my desires should constitute my will, while never complete, is one in which we can and constantly do readily engage. (That’s why the authority of higher-order volitions is neither mysterious nor arbitrary.)\textsuperscript{188}

Lenman seems to be making the following two points, among others: that we should not be surprised by the existence of what I am calling hefty higher-order attitudes, because we need them to keep our unruly and contrary desires in order; and that getting them in order is the central aim of moral inquiry. It isn’t very clear here, or in the surrounding discussion, whether the first-order desires he talks of are moral judgements or just desires of any old kind, nor whether the higher-order ‘commitments’ are meant to be moral judgements or something more analogous to epistemic norms. In another work, Lenman writes of accepting ‘relatively high-order moral and rational norms … governing the choice and revision of relatively low-order moral or rational norms’,\textsuperscript{189} which implies that moral judgements themselves admit of variations in ‘order’. By holding that higher-order norms may be norms of morality or rationality, Lenman seemingly combines elements of Blackburn’s and Gibbard’s accounts.

As before, these proposals perhaps do not definitely entail Heft, but would have odd implications if Heft were not true. Blackburn, Gibbard, and Lenman would be left with views on which a person’s judgements about which moral judgements it was moral and/or rational for her to hold under her circumstances would have no effect on which moral judgements she actually held. For these theorists, non-deductive moral reasoning, setting aside the changes of mind that it sometimes brings about, seemingly consists in measuring our moral attitudes (or our moral character, insofar as it is determined by those attitudes) against such higher-order judgements. So unless higher-order attitudes have heft, moral reasoning will be ineffectual.

\textsuperscript{188} Lenman, ‘What is Moral Inquiry?’, 78.
7.2 Impediments to Abandoning a Commitment to Heft

Blackburn’s early account of inconsistency among moral judgements as involving a mismatch between higher- and lower-order non-cognitive attitudes has been generally abandoned.\(^{190}\) Jeremy Schwartz and Christopher Hom seem to have shown how expressivists can account for the various ways of negating wrongness judgements without positing that they are attitudes of \textit{being for blaming for}, or any similar construction involving higher-order attitudes.\(^{191}\) If expressivists had an account of substantive moral theorizing that didn’t rely on Heft, they would be well on their way to unshackling themselves from Heft entirely. What are the obstacles to providing such an account?

The difficulty can be summarized in the following three points. First, substantive moral reasoning is the sort of reasoning whereby we arrive at, and change, our \textit{basic} moral judgements. By ‘basic moral judgements’ I mean moral judgements that are not derived. A person S has a \textit{derived} moral judgement that an action of type F has moral status W just in case S holds this judgement because she judges that actions of type G are W, and has a contingent belief that F actions are G actions.

Second, non-cognitivists (and hence expressivists) have good reason to accept the thesis that basic moral judgements are basic non-cognitive attitudes. By ‘basic non-cognitive attitude’, I mean a non-cognitive attitude that is not derived, i.e., that does not depend (for its rationality, if it is rational) on its possessor’s beliefs. Call the thesis that basic moral judgements are basic non-cognitive attitudes the Basic Judgement-Basic Attitude Thesis. I don’t believe I have come across an instance of an expressivist explicitly committing herself to this thesis. But I assume that they do, and it makes sense that they should, because doing so allows them to explain why people’s basic moral judgements tend not to change as a result of changes in their contingent beliefs. Note that, given that substantive moral reasoning is reasoning that produces changes in our basic moral judgements, the Basic Judgement-Basic Attitude thesis has the implication that substantive moral reasoning produces changes in our basic non-cognitive attitudes. It is reasoning that can create, alter, and kill off fundamental motivations.

\(^{190}\) Schroeder, \textit{Noncognitivism in Ethics}, 122; \textit{Being For}, 10n.3.  
\(^{191}\) Schwartz and Hom, ‘Why the Negation Problem’.
This brings us to our third point: some precisification of Heft, in conjunction with the hypothesis that we have higher-order attitudes concerning which non-cognitive moral attitudes we should have, seems to be the best explanation available to non-cognitivists of how substantive moral reasoning could change moralists’ basic non-cognitive attitudes. (By ‘moralists’ I mean people who make moral judgements.) After all, the non-cognitivist is hardly in a position to maintain that our *beliefs* about, say, the ‘coherence’ (in the ‘systematicity’ sense rather than the ‘rational consistency’ sense) of our basic moral judgements, are by themselves capable of reliably changing those basic moral judgements. If non-cognitivists conceded that beliefs had the power to reliably change basic non-cognitive attitudes, they would risk granting to their cognitivist rivals just what the latter needed to explain why moral judgements motivate.

7.3 *Problems with Heft and the Theory of Moral Reasoning it Supports*

Expressivists’ reasons for committing to Heft, then, are not frivolous. But the commitment is costly, and the account of substantive moral reasoning that becomes available upon committing to it is not altogether satisfactory. I will point out a strange implication of this account before considering some problems of the thesis itself. This implication is mainly owing to the Basic Judgement-Basic Attitude Thesis (which my theory of moral psychology denies). Observe that it is quite possible for a moral theorist to waver between accepting act utilitarianism and rule utilitarianism, now accepting one theory, now the other, in a fairly short space of time (as she reads the arguments for and against these theories, perhaps). Act and rule utilitarianism are comprehensive theories which make vastly diverging prescriptions. If the Basic Judgement-Basic Attitude Thesis is true, such a reasoner is undergoing an almost wholesale change in her basic moral motivations with each oscillation. This at least seems odd. A person’s basic motivational states are plausibly among the things that ‘make her who she is’; so if she wavers between basic moral judgements, and the Basic Judgement-Basic Attitude Thesis is true, it is almost as though her identity were in flux. But intuitively a change of mind between comprehensive theories need not be accompanied by a change in identity.

Admittedly, a person’s fundamental moral judgements also plausibly reflect ‘who she is’—a Kantian who thought it better that the heavens should fall than that a single

---

192 Blackburn might simply deny this, since he thinks that ‘there is a speed limit on change of values.’ *Ruling Passions*, 67. But it seems plausible that a compelling argument can change a person’s moral beliefs fairly quickly.
injustice should be done plausibly couldn’t become a utilitarian without becoming a different person. But act and rule utilitarianism are theories that appeal to people with the same sorts of concerns and preferences (a concern for welfare, a relative lack of concern for desert, a preference for simplicity and systematicity). They thus appeal to the same sorts of person. Intuitively and phenomenologically, the change of mind involved in wavering between these theories is more superficial, more akin to a change in descriptive beliefs. These appearances are at odds with the view that a change of basic moral judgements always involves a change of basic non-cognitive attitudes.193

My proposed moral psychology has the resources to explain the difference in our intuitions about the psychologies of the moral reasoner who sways between act and rule utilitarianism, and the moral reasoner who sways between Kantian deontology and act utilitarianism. It posits that the non-cognitive attitudes involved in basic moral judgements are preferences derived largely from stable basic altruistic desires, in conjunction with descriptive beliefs of a largely a priori nature. Changes in these beliefs might lead a reasoner to rationally change her preferences, but her basic desires constrain the sorts of preferences she could rationally arrive at. A radical change of basic judgements, such as we imagined the Kantian undergoing, would surprise us, because it would indicate either irrationality or a change in her (partially identity-constituting) basic desires.

What are the problems of Heft itself? Let us start by observing that it is not true in general that if I have a higher-order non-cognitive attitude my first-order attitudes will change in such a way as to satisfy it (or fulfil it, or realize it, etc.). I desire not to desire to eat salty food, and yet the latter desire persists. I may continue to add salt to my cooking, or I may not, but if I stop, it is against the urgings of the desire to eat salty food, and not because that desire has vanished. Not only is Heft not true in general, but I cannot think of any instance from my own experience that would clearly support it. By contrast, I can think of instances, even within the last few hours, when my beliefs changed in response to my evidence, or my derived desires changed in response to changes in my beliefs. So, if phenomenology and introspection provide good evidence

193 A non-cognitivist can apparently avoid the force of this objection by positing that moral judgements are intentions or planning states rather than more obviously desire-like attitudes. For, people’s intentions do seem to be very mutable, and can seemingly undergo dramatic changes without any attendant changes in the identities of their possessors. But the view that moral judgements are intentions risks proving too much: given the fickleness of intentions, why would we be surprised if the Kantian I described turned utilitarian?
as to the existence or non-existence of desires, the mental occurrences described by Heft must be at best exotic and elusive. Changes of mind about matters of basic moral principle, though much less common than changes in descriptive belief, do not seem to be nearly as rare as instances of higher-order attitudes spawning and killing off lower-order attitudes. So Heft is lacking in support from experience.

I admit that there is some intuitive plausibility in the idea that we can willfully repress, stifle, or drive from our thoughts desires that we would rather not have. This idea may also enjoy the support of phenomenology and introspection. But the plausibility of this idea gives no succour to Heft, which says that higher-order attitudes destroy rather than merely quell or silence non-cognitive attitudes. Nor does the idea suggest an alternative to Heft that would support expressivist theories as they are currently formulated. Existing expressivist theories postulate that moral judgements are non-cognitive attitudes, not that they are unrepresed non-cognitive attitudes. This postulation could be altered of course. But even if it were, the claim that when we come, though substantive moral reasoning, to reject an old basic moral judgement, we are merely driving from our thoughts an attitude that we still hold, and that is liable to resurface and spring back into action if our resistance slackens, is not very plausible.\textsuperscript{194}

It is possible that all we are really doing when we repress a desire—assuming we ever do—is effortfully focusing our attention away from those facts or stimuli in response to which it would move us to act. The idea that we are playing similar tricks on ourselves when we reject a basic moral judgement is likewise not very plausible.

Let us turn to the question of whether a good explanation of Heft can be supplied, given that it is true. Partisans of Heft might claim that its truth is explained by the fact that having a frustrated higher order attitude is rationally incoherent, and that we have a psychological faculty for ridding ourselves of rational incoherence. I cannot see any intuitive plausibility in the idea that having a frustrated higher-order attitude is rationally incoherent. It is not true in general that having a frustrated non-cognitive attitude is rationally incoherent—what makes higher-order attitudes different? And note

\textsuperscript{194} At least, it is not plausible if the quelled attitude was a highly theoretical moral judgement, like a judgement that rule utilitarianism is true. It is more plausible if the attitude was a judgement of a strongly intuitive (but perhaps theoretically suspect) sort, like the (pro-moral-luck) judgement that, among equally risky actions, an action is worse or more blameworthy if it happens to cause actual harm. But there is no reason why we should expect quelled attitudes to be exclusively non-theoretical judgements; theoretical moral judgements can obviously be rejected.
the following consequence of the idea, as it applies to desires at least. Any desire I have and can satisfy would be better served by the absence from my motivational set of competing desires. So, whenever I desire to φ and can φ, I will, if I am rational, have a derived second-order desire not to desire not to φ. If I am rationally required not to have frustrated second-order desires, then coherence will require me not to desire to not φ. Generalizing, coherence will, roughly speaking, require me not to have any opposing desires among options that are available to me. This would be surprising.

A similar observation actually supports the thesis that there is a kind of higher-order attitude such that one cannot rationally hold that attitude unless it is satisfied, at least in a wide range of circumstances. Alejandro Perez Carballo has pointed out that a rational agent who prefers to perform an action φ rather than not-φ, and knows that she will φ just in case she prefers to, will prefer that she prefers φ to not-φ, rather than that she prefers not-φ to φ, because this will mean that she φs. Thus, given that an agent knows, for a range of choice situations, that she will perform any given action just in case she prefers to, rationality requires her not to have frustrated second-order preferences, as long as those preferences are preferences among options in those choice situations.

Observe, however, that this does not provide any support for Heft. Heft posits that higher-order attitudes can, and reliably do, effect changes in their possessors’ lower-order attitudes. If Carballo’s argument supported Heft, it would imply that having a (second-order) preference to prefer φ would make it the case that one was rationally required to have a (first-order) preference to φ (in relevant cases). In fact Carballo’s argument implies no such thing, and it is natural to interpret it as implying that having a (first-order) preference to φ would make it the case that one was rationally required to have a (second-order) preference for preferring φ (in relevant cases). The rationally required second-order preference is one that is explained by and derived from the first-order preference. It has no heft in determining whether the agent has the first-order preference she must have if it is to be satisfied.

Expressivists could claim that Heft is explained by a requirement to satisfy one’s higher-order attitudes that is not a mere requirement of coherence, but which qualifies as a rational requirement on a more expansive conception of rationality. A prima facie

---

195 This is only a crude summary of Carballo’s argument. See Carballo, ‘Rationality & Second-Order Preferences’, esp. 4-6.
problem with this suggestion is that judgements of rationality that go beyond judgements of mere coherence seem to be straightforwardly normative; presumably expressivists will want to give them an expressivist construal. This being so, the capacity of the proposed rational requirement to explain our purported tendency to satisfy our higher-order attitudes itself stands in urgent need of explanation. To posit that we have a mental (or neurological) faculty for ridding ourselves of irrationality of this more expansive sort, and hence that we can discover the workings of this faculty through normative theorizing (construed in non-cognitivist fashion), would certainly be audacious.

Expressivists could maintain that Heft is not explained by any kind of rational requirement. Rather, they could just posit that we have a psychological or neurological faculty of some sort that, most of the time, brings it about that our higher-order attitudes are satisfied. A problem with this hypothesis is that moral reasoning seems to be a rationally evaluable activity. If I judge that the case for utilitarianism is stronger than the case against utilitarianism, I seem to be rationally required to form a utilitarian moral judgement—some variation on the judgement that an action is wrong just in case it fails to maximize happiness. Failing to do so would seem to be irrational in something like the way in which failing to proportion my descriptive beliefs to my evidence is irrational. But if the expressivist maintains that we have no rational requirement to satisfy those of our higher-order non-cognitive attitudes that normally explain the changes in our basic moral judgements that result from substantive moral reasoning, she cannot explain this appearance.

Expressivists avoid this problem if they extend their expressivist analyses of normative judgements to judgements of rational coherence, as well as judgements of rationality in its more expansive and obviously normative senses. Such expressivists, qua metaethicists, do not need to explain why it is incoherent to judge that the case for utilitarianism is stronger than the case against utilitarianism and yet reject utilitarianism, for they will take this to be a first-order normative question rather than a metaethical one. But there are metaethical questions in the vicinity to which we can demand that an expressivist of this stripe should provide answers. Why does the described mental state have roughly the same non-normative quality of dissonance and unresolvedness that we find in the mental state of a person who judges that the case against Francis Bacon’s authorship of *Hamlet* outweighs the case for it, yet remains a staunch adherent of the
Baconian theory? Why are people so apt to judge that the former mental state is irrational, in a suitably narrow sense?

The Heft-committed expressivist who takes coherence to be a first-order normative matter will, when trying to answer these questions, face difficulties similar to those facing expressivists who seek to actually establish the incoherence of having unsatisfied higher-order attitudes. For instance, if I desire to eat salty food despite desiring to desire not to, why does my mental state not seem dissonant in the same way that the Baconian theorist’s does?

7.4 Fiduciary Non-cognitivism’s Divorce from Heft

Fiduciary non-cognitivism dispenses with the need for Heft because, as I explained in the last chapter, it views substantive moral reasoning as a tightly constrained kind of instrumental reasoning. If a moralist’s motivations change as a result of substantive moral reasoning about matters of basic principle, this will usually be because (i) her voiceable preferences will have changed, typically as a result of changes in her instrumental reasoning or in her a priori beliefs about the implications of following relevant norms; and (ii) her discretional desires will urge her to act differently than before, because they will move her to act on her new voiceable preferences.

Fiduciary non-cognitivism avoids a commitment to Heft because it rejects the Basic Judgement-Basic Attitudes thesis. I observed in Section 7.2 that non-cognitivists had the following reason to accept this thesis: it allows them to explain why people’s basic moral judgements tend not to change as a result of changes in their contingent beliefs. But fiduciary non-cognitivism explains this too while rejecting the thesis, by positing that all propositions a consideration of which might lead to such variations in basic judgement are fundamentally inadmissible.

Fiduciary non-cognitivists can realistically aspire to free themselves from most commitments concerning rational choice, rational motivation, and the rationality of desires, beyond those posited by decision theory and austere versions of the Humean Theory of Motivation. This is in large part because they can realistically aspire to explain substantive moral reasoning as a species of instrumental reasoning.

196 They are, however, committed to the existence of indeterminate utilities (see 6.4.1 and 8.2), and the hypothesis that infinitesimal (or at least ‘near-infinitesimal’) credences exist and have significance in rational decision-making (3.3.3).
Moral Uncertainty

The phenomenon of uncertainty about matters of basic moral principle raises a metaethical puzzle. I shall argue here that fiduciary non-cognitivism has a more satisfactory solution to this puzzle than certain cognitivist views, specifically those that hold that we typically do not need an ulterior motive to act on our basic moral judgements. I describe the puzzle I have in mind in Section 8.1. Section 8.2 presents fiduciary non-cognitivism’s account of moral uncertainty. Section 8.3 explains how fiduciary non-cognitivism solves the puzzle.

8.1 A Puzzle About Acting Under Moral Uncertainty

The conjunction of the following two appearances is difficult to explain. First, rational people often seem to act on basic moral principles which they hold with imperfect certainty. (I should stress here, more emphatically than usual, that I have in mind rationality in the sense of having attitudes that are broadly coherent, or attitudes that are performing their functional roles properly, and not some more richly normative sense.) It is not uncommon for rational people to act on their moral judgements. But uncertainty about basic moral principles, in some degree, is surely ubiquitous among rational people; few if any could avow basic moral principles with the same confidence with which they asserted propositions about very simple arithmetic, for instance. So it is plausible that rational people not uncommonly act on imperfectly certain judgements of basic moral principle. Second, judgements of basic moral principle seem to be the upshot of reasoning about final ends, not instrumental reasoning. They seem to be or to define some subset of our ultimate concerns; and typically rational people don’t seem to need any ulterior motive to act on them.
It is hard to account for these appearances because it is hard to explain why it can be rational to act on judgements of basic moral principle if those judgements are not completely certain, and the actions do not serve some ulterior motive of the agent. It seems that we can sensibly ask the performer of such an action (call it φ) why she performed it given that she was, say, only ninety-nine per cent certain that the principle she acted on was true, and yet it is not clear what she can say in response. It is natural to suppose that she might say something like, ‘I wanted to do the right thing, and φ-ing seemed most likely to be right.’ But this would be at odds with the stipulation that she did not perform the action out of ulterior motives—it would suggest that she was acting to serve a further end of doing the right thing de dicto. What else could she say though? Perhaps she does not need to say anything. Perhaps the question is confused, like asking someone why she chose an option given that she preferred it. It doesn’t sound like a confused question though. After all, it wouldn’t make much sense for her to act on a moral judgement that she thought was false, and she isn’t completely sure that her judgement isn’t false, and yet she is acting on it anyway, without an ulterior motive—this seems to require explanation. And suppose that the agent answered the question thus: ‘I didn’t need a reason for acting on my highly confident though not- quite-certain moral judgement. I just did it.’ This statement would hardly assure us of the agent’s rationality if it had not been established by stipulation.

Of the two of appearances that I cited above, the most doubtful is the appearance that judgements of basic moral principle are the upshot of reasoning about final ends, and that we typically act on them as such, i.e., for non-instrumental reasons. This is unobvious to many philosophers. But the explanation for the rationality of acting on uncertain judgements of basic principle that suggests itself if we reject this appearance faces a challenge of its own concerning action under moral uncertainty. The explanation I have in mind is as follows: we have a standing desire to do the right thing, whatever that may turn out to be. This motivates us, as we would expect, to choose actions that we judge to be permissible and eschew actions that we judge to be impermissible, and motivates us more strongly the more confident our judgements of permissibility and impermissibility are. Call this the Instrumental Explanation. A consideration of certain

\[197\] A desire to do the right thing in the de dicto sense is a general desire to do the right thing, whatever that may turn out to be. By contrast, a desire to do the right thing in the de re sense is an underived desire, for each thing that is in fact non-derivatively right, to do that thing. Cf. James Dreier, ‘Dispositions and Fetishes: Externalist Models of Moral Motivation’, Philosophy and Phenomenological Research 61, no. 3 (2000): 621-2.
kinds of cases suggests that this explanation either predicts that moralists will behave in ways that are intuitively surprising, or it posits that they are far less concerned to do the right thing than we would intuitively expect.

Suppose that a woman is considering having an abortion. She wants to have an abortion, for intuitively non-frivolous reasons. She is highly confident that abortion is permissible, confident enough to avow that this is so, but has a small degree of confidence—say, one per cent—that abortion is impermissible, and indeed morally indistinguishable from killing an innocent adult. She is virtually certain that not having an abortion is permissible. It seems that if the instrumentalist explanation is true, we should expect her not to have the abortion. After all, from her point of view the moral desirability of having an abortion ought to be similar to that of demolishing a building (for similarly non-frivolous reasons) when one has a one per cent credence that a human being is inside. But I take it that this is not how we expect her to behave given the state of her moral judgements.

If someone judges that abortion is permissible, we expect her to have an abortion if she wants one, even if she is not completely certain that abortion is not gravely wrong. A woman who wants to have an abortion, who agonizes about whether or not abortion is wrong, and finally decides that it is not, will typically take this to settle the matter of whether she will have an abortion. She will not continue to forbear from doing so, ‘just to be on the safe side’, because she thinks that there is a hundredth chance that she is mistaken. A defender of the Instrumental Explanation could maintain that her theory does not predict that the woman will forbear from having an abortion by hypothesizing that people’s de dicto desires to do the right thing are not normally strong enough to induce them to avoid actions because they have such small probabilities of being immoral. But if this is so, why are we not surprised by the strength of a person’s moral motivations if she refrains from demolishing a building because she has a 0.01 credence that someone is inside it, and she judges that it is wrong to kill?

I think that this problem is surmountable. A defender of the Instrumental Explanation can avoid making false predictions in the abortion case by giving an

---


199 This example is adapted from Alexander A. Guerrero, ‘Don’t Know, Don’t Kill’, Philosophical Studies 136 (2007): 68.
account of permission judgements and moral motivation that is similar in important respects to the fiduciary non-cognitivist account. I shall discuss this further when I have explained how fiduciary non-cognitivism approaches the case in Section 8.3.

Cognitivists who reject the Instrumental Explanation will probably want to posit requirements of rationality that explain why rational agents sometimes act on imperfectly certain judgements of moral principle without ulterior motives. These requirements must, of course, be the sorts of rational requirements whose violation would indicate a failure of coherence, or a failure of the agent’s attitudes to perform their functional roles. This is, after all, the sort of rationality that I was talking about when I described our explanandum, and these are the sorts of rational requirements that can most plausibly figure in an explanation of thought and behaviour. Requirements of rational coherence, or of the functional role of attitudes, are ones that we expect (in the descriptive sense) all agents to be influenced by to some degree. By contrast, principles of rationality of a more ‘expansive’ and straightforwardly normative sort are ones that we will only expect agents to be guided by if they happen to accept those principles, and we will not normally be justified in assuming that rational agents accept them.

A few philosophers have proposed principles of rationality governing actions under uncertainty about basic moral principles, and normative principles more broadly. But these principles of rationality tend to be quite normative and revisionary, and are thus ill-suited to an explanation of the tendency of ordinary rational agents to act on imperfectly certain judgements of basic moral principle. It seems unlikely that they are proposed as requirements of rationality in the ‘coherence’ or ‘functional role’ senses. They tend to prescribe maximizing expected conformity to the true basic normative requirements (think of ‘conformity’ here as a graded notion: violating weighty principles is a greater failure of conformity than violating less weighty principles), or some variation on this idea. Such ‘maximizing’ proposals tend to require the woman in the abortion case to forbear from having an abortion just as the


201 Proposals include ‘maximizing expected objective value’ (Sepielli, ‘What to Do’, 11-12), ‘maximizing expected choiceworthiness’ (Wedgwood, ‘Akrasia and Uncertainty’, esp. 493-497), and ‘maximizing expected degree of moral rightness’ (Lockhart, Moral Uncertainty, e.g. his ‘PR4’, 82).
Instrumental Explanation does, at least assuming that moral principles have the authority in decision-making that they seem to have. That moralists should maximize expected conformity to the true moral requirements may or may not be a good normative principle, but it is too revisionary to figure in an explanation of ordinary moral thought and talk. It is unlikely that the folk tacitly accept it, and that this acceptance explains why they act on uncertain judgements of basic moral principle and expect others to do the same.

Suppose that a hypothesized principle of rationality was devised in such a way as to ensure that it prescribed acting on uncertain judgements of basic moral principle in all and only those circumstances in which the folk expect people to act on such judgements. And suppose that the explicitly hypothesized principle was a principle of rationality of a sort that is relevant to my argument (e.g. a principle of coherence). This sort of explanation would still be inadequate unless a plausible argument could be supplied as to why it was a requirement of coherence or of the proper functioning of attitudes.

Michael Smith is the only philosopher I am aware of who rejects the Instrumental Explanation and who has explicitly posited a requirement of coherence that is relevant to an explanation of commonsense expectations about how people will act under uncertainty about basic moral principles. The requirement does not directly concern which actions one should perform or intend under such uncertainty. Rather it concerns which basic desires it is rational to have under such uncertainty. Uncertain evaluative judgements cause rational people to desire to act on them, though the desires will be weaker the more uncertain the judgements are.202 This explains why rational people act on uncertain judgements of basic principle.

According to Smith desirability judgements are judgements about what we would advise ourselves to do if our desires were better informed, more coherent, and more unified. Coherence requires you to desire to act on your desirability judgements because to do otherwise is to have desires ‘that you yourself disown’.203 And the strength of the desires coherence requires you to have is proportionate to the confidence of your judgements; Smith thinks that a comparison of cases in which the strengths of your desires are and are not proportionate in this way could be described which would

203 Smith, ‘Evaluation, Uncertainty, and Motivation’, 311-12 (quote is on 312).
intuitively establish this.\textsuperscript{204} I lack the intuitions about coherence that Smith appeals to here, but I’m sure some readers of Smith will respond differently. Some sort of argument against Smith is thus in order. I’ll argue that Smith’s theory fails to explain our expectation that the woman in the abortion case would have an abortion upon deciding that this was permissible.

Smith thinks, roughly, that a coherent moralist’s motivation to act on a hypothesized moral requirement will vary directly with her understanding of the \textit{importance} of that hypothesized requirement, if her confidence in it is positive and real-valued.\textsuperscript{205} And he claims that his theory predicts this.\textsuperscript{206} So it seems at least possible that his theory predicts that the woman in the abortion case who judges that she may have the abortion would, assuming she is coherent, be strongly motivated to avoid doing so even though her confidence that abortion is wrong is very low. After all, she judges that the hypothetical requirement to forbear from abortion, if it exists, is very important. (I assume that her high confidence that abortion is morally optional would not make it rational for her to conceive a basic desire of any strength to \textit{have} an abortion.)

Perhaps Smith’s theory doesn’t predict this. When reading Smith, I have let myself be guided by the idea that rational requirements governing basic desire in cases of moral uncertainty are supposed to be analogous to rational requirements governing derived desire in cases of descriptive uncertainty. That is, we are to ‘proportion’ our basic desires to our ‘expected importance’ assessments in much the same way in which we are to proportion our derived desires to our expected utility assessments in instrumental reasoning. But perhaps this analogy is misleading; perhaps the relationship between rational motivation and ‘expected importance’ assessments is not as straightforward as the analogy would suggest. I suppose that the way to settle the matter is to consult once again our intuitions about coherence. Here I am at a disadvantage, for, as I have said, I lack Smith-friendly intuitions—I don’t see anything incoherent about lacking a basic (\textit{de re}) desire to do the thing that you would advise yourself to do if your desires were more ‘coherent’ and ‘unified’, even absent uncertainty about the hypothetical advice. So it is hard for me to know which direction Smith-friendly

\textsuperscript{204} Smith, ‘Evaluation, Uncertainty, and Motivation’, 315.
\textsuperscript{205} See Smith, ‘Evaluation, Uncertainty, and Motivation’, 309. I am taking some liberties in this interpretation, for Smith only talks about value judgements, not judgements of requirement and permissibility.
\textsuperscript{206} Smith, ‘Evaluation, Uncertainty, and Motivation’, 314. The qualifications from my previous footnote apply here also.
intuitions point in the abortion case. But insofar as Smith-friendly intuitions are explained by the principle that it is incoherent to have desires ‘that you yourself disown’, they should not yield any clear verdict in the abortion case. It is simply not clear what desires the woman in the abortion case ‘disowns’. So I tentatively conclude that Smith’s theory makes no determinate prescription about what the woman ought to do, and so fails to explain our expectation that she will go through with the abortion.

8.2 Moral Uncertainty for Fiduciary Non-cognitivists

8.2.1 A Refinement of My Account of Entrusted Parties’ Obligations and Entitlements

In Subsection 8.2.2 I will set out fiduciary non-cognitivism’s account of uncertainty about basic moral principles and its consequences. But first I need to revise my position on what, as a general matter, entrusted parties must do to discharge their obligations to accommodate their partners’ weighted desires, and what they must do to avoid exceeding their discretion when they choose not to accommodate their partners’ weighted desires. (The revision is not very relevant to 8.2.2, but significantly affects my claims in Section 8.3.) The revision is motivated chiefly by intuitions about the propriety of acting on voiced or voiceable preferences that are irrational or based on false beliefs. Briefly, if an entrusted party acted on his voiceable preference in defiance of his partner’s, it does not seem that he would strain her trust if did so knowing that her preference depended on false beliefs or poor reasoning. And it does seem that an entrusted party might seriously strain his partner’s trust if he resisted her preference to act on a very poorly reasoned voiceable preference of his own. The revision aims to accommodate these appearances. It hypothesizes a potential disparity between the things an entrusted party must do to discharge his ‘entrusted concern’ to accommodate his partner’s weighted desires, and the things his partner voiceably prefers that he should do. And it hypothesizes a potential disparity between the things an entrusted party must do to avoid exceeding his discretion, and the things he voiceably prefers to do.

I call the conjunction of the hypotheses that I make here the ‘Disparity Hypothesis’, in contrast to the simple ‘Parity Hypothesis’ whose truth I have assumed so far. The Parity Hypothesis is the view that entrusted parties have discretion to act on their voiceable preferences, and are obliged to accommodate their partners’ voiceable
preferences, whether or not those preferences are rational and well-informed (though acting on irrational discretion-licencing preferences or sincerely voicing irrational preferences will presumably involve some sort of conventional impropriety and often harm one’s reputation).

The Disparity Hypothesis is my more considered view of the nature of entrusted parties’ obligations and entitlements. I think that the Parity Hypothesis, though incorrect, is accurate enough for many purposes, which is one of the reasons I have delayed introducing the Disparity Hypothesis until now. Considering its subject matter, the material of this subsection belongs in Part I—perhaps even in Chapter 1. But I was reluctant to include it earlier because it would not have had a great bearing on any of my arguments in Part I, and because doing so would have tested the patience of readers who were concerned to see some metaethical payoff from my theories about trust relationships and trust networks. Taking account of the Disparity Hypothesis consistently throughout this thesis would have required correcting some of my arguments and hypotheses in Chapter 3. But I think that, given Chapter 3’s complexity, it was acceptable to assume the truth of the simpler Parity Hypothesis in that chapter, and that this was in keeping with the other simplifying assumptions I made there (specifically in 3.3.3).

The Disparity Hypothesis will employ a few variations on the idea of ‘expected weighted utility for a person’. The expected weighted utility of \( \phi \) for \( S \) by \( T \)'s lights is—in a sense I will need to clarify—\( T \)'s assessment of the summed probability-weighted utilities of the outcomes of \( \phi \), where utilities are understood in terms of the satisfaction of \( S \)'s weighted desires, and probabilities of outcomes are determined by lines of epistemic support that take into account only propositions that are admissible for \( S \) (or as I shall say, these probabilities are ‘constrained by \( S \)'s admissible ground-types’). The expected weighted utility of \( \phi \) for \( S \) by \( T \)'s lights is a value describing or representing an actual mental state—a rationally evaluable mental state belonging to \( T \). And it will be a different kind of mental state depending on whether or not \( S \) and \( T \) are the same person.

If \( S \) and \( T \) are different people, the expected weighted utility of \( \phi \) for \( S \) by \( T \)'s lights will represent a cognitive state. It will be something like \( T \)'s belief about how ‘voiceably preferable’ \( \phi \) ought to be for \( S \), given \( S \)'s basic desires, and in light of the actual facts (not \( S \)'s beliefs). More precisely, its value will be the \( X \) such that \( T \) believes that the summed probability-weighted utilities of the outcomes of \( \phi \), where utilities are
understood in terms of the satisfaction of S’s weighted desires, and probabilities are constrained by S’s admissible ground-types, is most likely to be X. This belief will typically be based on T’s beliefs about the content and strength of S’s weighted desires, as well as his credences concerning the various possible states of the world conditional on φ, ignoring any facts that fall under ground-types that are inadmissible for S. (It might not though—S might have arrived at it by some more direct means, e.g., through testimony, without the aid of any psychological theorizing on his part.)

By contrast, the expected weighted utility of φ for T by T’s lights will represent a non-cognitive state broadly reflecting T’s admissible credences in the possible outcomes of φ and the weighted utilities for T of those outcomes. Loosely, it represents the degree to which T voiceably prefers φ. This mental state is rationally evaluable, given T’s other attitudes, by its conformity to the prescriptions of decision theory—or rather, to a circumscribed kind of decision theory in which the probabilities involved are T’s admissible credences and the utilities are understood in terms of the satisfaction of T’s weighted desires rather than all-things-considered desire satisfaction. The option with the highest expected weighted utility for T by T’s lights is necessarily the option for which T has a voiceable preference. I shall take this to be analytic.207 When, in previous chapters, I talked about a person’s ‘expected weighted utilities’, I always meant by this the expected weighted utilities for her by her lights.

To reiterate, the expected weighted utility of φ for T by T’s lights is not a belief. It is, for all I shall say, possible for there to be an expected weighted utility of φ for T by T’s lights, with some value X, even if (i) T has no beliefs about the summed probability-weighted utilities of φ’s outcomes, where utilities are understood in terms of the satisfaction of T’s weighted desires, and probabilities are constrained by T’s admissible desire-types; or (ii) T rationally believes that the sum of these probability-weighted utilities has some value other than X.

The final idea I need to introduce is that of the expected weighted utility of φ for a person S relative to a body of evidence E. This is the value X such that, according to E, the summed probability-weighted utilities of the outcomes of φ, where utility is understood in terms of S’s weighted desire satisfaction, and probabilities are constrained

207 I am admittedly unsure whether it is better to treat this as analytic or as a requirement of coherence. Potentially at stake here is the success of our explanation for the possibility of having irrational (e.g. cyclical) combinations of voiceable preferences, and hence, perhaps, the possibility of holding inconsistent judgements of basic moral principle.
by S’s admissible ground-types, is most likely to be X. It is equivalent to the expected weighted utility of φ for S by the lights of a hypothetical person (not S) who is perfectly rational and logically omniscient, and whose evidence is E. ‘Evidence’ must be understood in an expansive way here, one which includes all the considerations and appearances that may epistemically support or undermine a belief, including such things as intuition and introspective appearances.

The Disparities Hypothesis makes two main claims. First, in general, an entrusted party B is obliged, not to accommodate his partner A’s preferences, but rather to perform the action with the highest expected weighted utility for A, relative to B’s evidence. ‘B’s evidence’ here includes B’s evidence about the content and strength of A’s weighted desires, as well as his evidence about the probabilities of outcomes of available options. Note that the highest expected weighted utility for A, relative to B’s evidence is not the same as the highest expected weighted utility for A by B’s lights. The latter represents an actual mental state of B, the former does not. We might say that the former represents a mental state that B would have, holding his evidence fixed, if he were perfectly rational and logically omniscient.

The preferences that A voices when she scrupulously makes demands of B will nonetheless, if they are rational, be based on A’s evidence, not on what A takes B’s evidence to be. A’s sincerely voiced preferences will be for the options that have the highest expected weighted utility for A by A’s lights, i.e., for her voiceably preferred options. It makes sense for A to voice such a preference, even though her preferred option may not be the option B is obliged to choose, because part of the purpose of A’s voicing the preferences she voices will normally be to give B evidence about what it would take to best satisfy her weighted desires. This will be evidence about the content and strength of her weighted desires, as well as about the probable outcomes of options.

Second, it is B’s weighted desires and evidence that determine whether he has discretion to resist A’s demands—or, more accurately, to resist performing the action.

---

208 I think that the terms of entrusted concern relationships must ordinarily include norms of some sort requiring the parties to cooperate in the epistemic project of determining what would best satisfy the entrusting party’s weighted desires (and perhaps also the entrusted party’s). Parties engage in verbal exchanges involving the voicing of opposing preferences, demanding and giving grounds for preferences, partly as part of this shared project. These norms do not derive from any trust-based obligations I have posited so far; rather, they are explained by the fact that it serves both parties’ interests for the entrusted party to have accurate relevant means-end beliefs.
that has the highest expected weighted utility for A relative to B’s evidence. B has discretion to φ in defiance of A’s demands only if φ has the *highest weighted utility for B, relative to B’s evidence*. This means that B only has discretion to act on his voiceable preferences when they the preferences he would have if he were rational: if his relevant preferences were coherently proportioned to his weighted desires and credences, and his relevant credences were coherently proportioned to his evidence. When B voices a preference to φ in defiance of A’s demands, part of his purpose will typically be to communicate that φ-ing has the highest expected weighted utility for him relative to his evidence, and hence that he has discretion to φ.

Why am I positing that rational expected weighted utilities relative to entrusted parties’ evidence, and not voiceable preferences, are fundamental to the determination of what entrusted parties are obliged and have discretion to do? As I have foreshadowed, some such postulation seems necessary to explain certain commonsense expectations about what would strain an entrusting party’s trust.

Imagine the following series of events involving the *dramatis personae* of *Football (Mother and Son)*: John’s mother sincerely voices a preference and explains why she has it; John later realizes that her preference depended on an error in her instrumental reasoning or an inaccuracy in her beliefs about the probabilities of outcomes; John does not accommodate the preference; John’s mother complains about his failure to accommodate the preference; John points out the error or inaccuracy; she acknowledges the error or inaccuracy. It seems to me that we would now expect John’s mother to withdraw her complaint entirely. But if John were obliged to accommodate his mother’s preferences, and not instead to do what, given his evidence, it would take to satisfy the weighted desires on which they are based, it would be harder to explain this. His failure to accommodate her ill-founded preference would have been a strain on her trust, and she might still have ‘held it against him’, so to speak. My first postulation avoids this odd implication.

Now imagine a different series of events. (Assume that John and his mother both have the same evidence relevant to John’s decision and that this evidence does not change throughout the story.) John acts on a preference of his own in defiance of his mother’s; his mother demands to know why he did not act on her preference; John

---

209 Which is not to say that she will not have some other complaint about his action, as she might if he didn’t do what by his lights would have had the highest expected weighted utility for her.
explains his preference to her; John’s mother points out an error in the reasoning behind the preference, revealing that, given his evidence, he should not have expected his action to serve his weighted desires as well as her preferred action would have. Intuitively, John’s choice to defy his mother’s preference so that he could act on a preference that was poorly thought through is the sort of thing that might strain his mother’s trust, especially if the error were the result of laziness in deliberation.

Properly accounting for these appearances would require some rather detailed theorizing. It might require, for instance, postulating that entrusted parties conventionally have ‘excuses’, of varying strength, for exceeding their discretion in some circumstances—for instance, if they acted on voiceable preferences that were irrational, but were the result of reasoning of that met certain standards. The strain an entrusted party placed on his partner’s trust by exceeding his discretion would vary inversely with the strength of his excuses.210 Such detailed theorizing is beyond the scope of this work. But the most rudimentary lesson of the example is one that my theory of entrusted concern relationships must account for: viz., if an entrusted party acts on his voiceable preference, and that preference is the result of poor reasoning, he can strain and perhaps violate trust. My second postulation—that it is the expected weighted utilities that it would be rational for an entrusted party to have given his evidence that determine what he has discretion to do, and not his voiceable preferences—allows me to accommodate this appearance.

I don’t think that this argument clearly shows that my second postulation is superior to the following rival hypothesis: an entrusted party B exceeds his discretion and strains his partner B’s trust if he acts on a voiceable preference in defiance of his obligation to accommodate A’s weighted desires, and that preference is not the voiceable preference B would have had if he were perfectly rational and well-informed. My postulation implies, by contrast, that an entrusted party does not strain trust if he acts on a preference that was rational given his evidence, even if doing so satisfied his weighted desires poorly relative to alternatives. The rival hypothesis seems simpler than my proposal: we can sum it up by saying that entrusted parties have discretion to do only what would best satisfy their weighted desires, without invoking the complicated idea of ‘expected weighted utilities for X relative to Y’s evidence’.

210 I have independent reason to posit the existence of excuses for failing to accommodate entrusting parties’ demands. I must do so if I am to plausibly explain, in some contexts, why trust relationships permitting third-party disagreement exist. See Appendices B and C.
However, I think it would be surprising if this rival hypothesis were correct. Entrusted concern relationships that worked in accordance with this rival proposal would generally serve parties’ interests less well than they might have. I presume that, if an entrusted party has discretion to do only what would best satisfy his weighted desires, his discretionals desires will urge him to either yield to his partner’s wishes, or choose the option with the greatest probability of maximally satisfying his weighted desires. But an entrusted party’s weighted desires would tend to be less well satisfied if, instead of aiming to do what had the highest expected weighted utility for him relative to his evidence, his practical reasoning were aimed at choosing the option with the greatest probability of maximally satisfying his weighted desires.

This is suggested by the consideration of a sort of example that will be familiar to moral philosophers. Frank Jackson famously describes the following case:

Jill is a physician who has to decide on the correct treatment for her patient, John, who has a minor but not trivial skin complaint. She has three drugs to choose from: drug A, drug B, and drug C. Careful consideration of the literature has led her to the following opinions. Drug A is very likely to relieve the condition but will not completely cure it. One of drugs B and C will completely cure the skin condition; the other though will kill the patient, and there is no way that she can tell which of the two is the perfect cure and which the killer drug.\textsuperscript{211}

Assume, as seems obvious, that the complete cure would be somewhat better than the incomplete cure, and that killing the patient would be disastrous. Jill will tend to do more good over time if, in choices like these, she chooses the option with the greatest expected value (in this case, drug A) than if she chooses the option with the greatest probability of realizing the best outcome (tied here between drug B or C). A policy of choosing the options with the greatest probability of realizing the best outcomes is insensitive to the risks that may come with those options.

An entrusted party will sometimes face choices analogous to Jill’s, in which the option that has the best chance of maximally satisfying his weighted desires is risky, and he would tend to better satisfy his weighted desires over time if his decision-making were sensitive to these risks. Thus, if it is to the advantage of parties to an entrusted concern relationship for the entrusted party to be allowed to better satisfy his weighted desires over time, all else being equal, then it would be surprising if entrusted parties

had and only had discretion to choose the options that best satisfied their weighted desires, instead of having and only having discretion to choose the options with the highest expected weighted utility for them, either by their lights, or relative to their evidence.

I think it is typically to the advantage of parties to entrusted concern relationships for the entrusted party to be allowed to better satisfy his weighted desires over time, all else being equal. An entrusted party’s discretion to act on his weighted desires is typically meant to make his commitment to accommodate his partner’s weighted desires less burdensome, in a way that is not too irksome to his partner. Given that an entrusted party is sometimes going to be defying his partner’s preferences, his partner will generally prefer that he should well satisfy his weighted desires when he does so, because these will be desires that she is comparatively willing that he should indulge. And the entrusted party will, on the whole, ‘want’ his weighted desires to be well-satisfied, because they are his desires. Thus we should expect that, at least in normal entrusted concern relationships, an entrusted party enjoys discretion to choose the option with the highest expected weighted utility for him, either by his lights, or relative to his evidence. Our intuition that entrusted parties can strain trust by acting on ill-considered voiceable preferences indicates the latter possibility.

This concludes my description and defence of the Disparity Hypothesis. It is not clear whether I should hold that the Disparity Hypothesis is necessarily true of all entrusted concern relationships, or whether I should hold that it is merely true of almost all actual ones. But for the sake of specificity, I’ll adopt the former position.

The hypothesis barely affects my account of the mental states involved in non-third-party robust non-cognitive disagreement. The hypothesis implies that an entrusting party A and her partner B, the fundamental determinants of which of B’s actions would strain or violate A’s trust are the expected weighted utilities for B, relative to B’s evidence, and the expected weighted utilities for A, relative to B’s evidence, of the options available to B. But the hypothesis does not imply that these are the things that put an entrusted party in robust non-cognitive disagreement with his partner. It is a clash of their voiceable preferences (rational or not) that puts two people

\[212\] It affects this account at all only insofar as it affects what things count as entrusted concern relationships, and hence what things count as voiceable preferences.
in robust disagreement, and not a mismatch among the options that, for each party, have the highest expected weighted utility for her relative to the entrusted party’s evidence.

Appearances seem to require this. Suppose that two people report mutually unsatisfiable preferences about what one of them is to do. Subsequent discussion, in which neither party provides the other with new information about the probable outcomes of available options or about her basic desires, reveals an error in the reasoning behind one party’s preferences. This discovery leads to a change in that party’s preferences, with the result that the two parties report mutually satisfiable preferences. We will presumably lack an intuition that they disagree after the change of preference, but the change of preference will not change our intuition that the parties robustly disagreed before the error was discovered, if we had such an intuition. And yet presumably the parties’ basic desires and evidence—including their evidence about each other’s basic desires—will not have changed, just the reasoning behind one party’s preferences. On the assumption that they disagreed in virtue of sharing an entrusted concern relationship, this means that the expected weighted utilities for the entrusted party, relative to his evidence, and the expected weighted utilities for the entrusting party, relative to the entrusted party’s evidence, will not have changed. These utilities supervene on the entrusted party’s weighted desires and evidence, and do not depend on the quality of anyone’s reasoning. Positing that the parties’ voiced preferences were what put them in disagreement, and not the utilities just described, is the most straightforward way to explain our intuitions about the parties’ disagreement before and after the change of preference.

As I have acknowledged, taking proper account of the Disparity Hypothesis would require amendments to several part of this dissertation. I will only address one of these. It seems more important that I should make the relevant corrections to my metaethical theory, and my claims about its predictions, than that I should revise my arguments for the general empirical theses concerning trust networks that I defended in Chapter 3. At present I am content to take the position that those claims were only a sketch for a rigorous account of the workings of dual-order fully connected networks and the reasons for their existence. But I do want to establish that my commitment to the Disparity Hypothesis does not greatly undermine my explanation for moral disagreement. I undertake this task in Appendix E.
Confidence and Indecision in Matters of Basic Moral Principle

What is fiduciary non-cognitivism’s account of uncertainty about basic moral principles, and the practical implications of such uncertainty? Regarding uncertain moral judgements whose uncertainty is owing to uncertainty about basic principles, I cannot give a full account of what mental states these uncertain judgements consist in, only an account of their non-cognitive components. There are two phenomena that I need to explain, which I will call ‘confidence’ and ‘indecision’.

**Confidence** is analogous to degree of cognitive belief. Loosely and intuitively, one’s confidence in a moral judgement is how likely one thinks it is to be true. A person’s confidence in a moral judgement may depend on her credences concerning descriptive facts, as well as on her state of mind regarding basic moral principles; I shall, of course, only be concerned with the latter. Here are a few platitudes about confidence that my account of it must save. A person can have degrees of confidence in moral judgements that she does not herself hold; indeed, she will typically have degrees of confidence in moral judgements that are incompatible with those she holds herself. Necessarily, she will have a higher degree of confidence in a moral judgement she holds than in any moral judgement that is inconsistent with it, if she is rational. A moralist’s degrees of confidence in moral judgements, insofar as they are explained by her state of mind regarding basic principles, normally depend greatly on the state of her moral reasoning.

**Indecision** is, intuitively, a state of uncertainty resulting in an absence of moral judgement. If a moralist is undecided about whether some action φ is wrong, she does not judge that φ-ing is wrong or permissible. If she must choose between φ-ing and not-φ-ing in the near future, her indecision will characteristically lead to inaction and prolonged deliberation, or, if prolonged deliberation is not possible, to a choice that she will make unhappily and reluctantly. People can apparently be undecided about the morality of actions as a result of uncertainty about contingent non-moral facts, and as a result of having a certain state of mind regarding basic moral principles; I am only concerned with indecision insofar as it is explained in the latter way. Like confidence, indecision (where it has this latter sort of explanation) generally depends on the state of one’s moral reasoning. Indecision is related to confidence in at least the following way: if a moralist is undecided about whether or not φ-ing is wrong, her confidence that φ-ing is wrong will be not be clearly higher or lower than her confidence that φ-ing is not wrong.
The general metaethical problem I described above—the problem of explaining why it is sometimes rational to act on basic moral principles of which you are not completely certain—is a problem related to confidence, not indecision. A person who is undecided about the morality of φ-ing is in no immediate danger of acting on an uncertain judgement that she ought to φ, for she has no such judgement.

Fiduciary non-cognitivism’s account of the non-cognitive attitudes involved in indecision about basic principles can be sketched quickly, so I shall carry out this task before turning my attention to confidence. Moral indecision involves norm-indecision. Recall my account of norm-indecision from 6.4.1: for a moralist S so be norm-undecided about φ-ing is to be such that (a) S’s expected weighted utilities fail to discriminate among two or more of S’s best preferred comprehensive norm-partitions, and (b) these best preferred comprehensive norm-partitions include norm-partitions both forbidding and requiring φ. (The Disparity Hypothesis doesn’t require any alteration of this account, beyond a clarification that ‘S’s expected weighted utilities’ are S’s expected weighted utilities for S by S’s lights.)

What follows is only an approximation of an account of indecision. A full statement of the necessary and sufficient conditions for indecision would need to take account of the weird possibility that an agent might be in different states of norm-indecision in the various roles she occupies in the first and second orders of the moral trust network. Such an account would be relatively lengthy and complicated, and would achieve a greater level of thoroughness than is called for at present.

Suppose that C is a set of circumstances. C may be dispositionally non-specific to any degree.

**Indecision about requirement.** A person S is in a state of indecision about whether or not φ-ing in C is necessarily required iff there is some C* such that C* is a dispositionally-specific circumstance that entails C, and S is norm-undecided between norms forbidding and norms requiring φ-ing in C*.

**Indecision about permissibility.** A person S is in a state of indecision about whether or not φ-ing in C is necessarily permissible iff there is some C* such that C* is a dispositionally-specific circumstance that entails C, and the agent in C* is pre-trust disposed to φ, and S is norm-undecided between norms forbidding and norms requiring φ-ing in C*.

This account of indecision, in conjunction with the rest of my theory of moral psychology, predicts the platitudes about indecision that I listed above. First, indecision results in an absence of moral judgement. Someone who is norm-undecided between norms requiring or forbidding φ-ing in a dispositionally specific circumstance C* lacks
a voiceable preference between conforming to norms requiring or forbidding φ-ing in C*. The expected weighted utilities for her, by her lights, fail to discriminate between these options because these utilities are indeterminate. This means that she lacks a voiceable preference for conformity to any norm-partition whose extension includes C*, and hence that she lacks any moral judgement such that holding that judgement involves having such a voiceable preference, including the judgements that φ-ing in C is wrong, required, or (if the agent in C* is pre-trust disposed to φ) permissible.

Second, indecision will characteristically lead to inaction and prolonged deliberation, or to a reluctant forced choice. The platitude I need to explain here is that an agent who expects to find herself in dispositionally non-specific circumstance C in the near future, and who is undecided about whether φ-ing in C is wrong or permissible, will prevaricate, engage in prolonged deliberation, and so on. There is a problem here: I can easily explain why someone would exhibit inaction etc. if she (i) expected to find herself in dispositionally specific circumstances C*, where C* entails C, and (ii) was undecided about whether or not φ-ing in C is wrong because she was undecided about whether or not φ-ing in C* is wrong. She lacks a discretion-licencing preference to φ or not-φ in C*, and so would violate trust qua moral entrusted party if she did so. So, as long as inaction is possible and by her lights permissible, we should expect her to delay action, and try her best to form a voiceable preference, to avoid such a violation.

But it is apparently possible for a moralist to (i) expect to find herself in dispositionally specific circumstance C*, where C* entails C, (ii) have a voiceable preference for following norms forbidding or requiring φ-ing in C, and (iii) be undecided about whether or not φ-ing in C is wrong because she is undecided about whether or not φ-ing in dispositionally specific circumstance C′ is wrong, where C′ entails C, and C* and C′ are different circumstances. In this case my theory does not predict inaction etc., because the moralist has a voiceable preference that is applicable to the circumstance she expects to find herself in.

However, we should expect this sort of situation to be very unusual, given the hypotheses about the nature of altruistic desires that I made in 4.1.4 with the aim of explaining the importance of permissibility judgements in light of my claim that they are reducible to requirement judgements. There I posited that the extent to which an

---

213 I touched upon this topic, with respect to norm-indecision, in 6.4.2.
action satisfies altruistic desires may often depend on whether the agent is pre-trust disposed to perform it, but will rarely depend other features of her dispositional circumstances (e.g. strength of pre-trust preferences, which specific sub-option she is pre-trust disposed to choose or perform). These further features are, by and large, not altruistically relevant. So, among those C-entailing dispositionally specific circumstances in which the agent is pre-trust disposed to φ, it will rarely be the case that a moralist will be undecided about whether or not φ-ing is wrong in some of those circumstances but not others. And among those C-entailing dispositionally specific circumstances in which the agent is pre-trust disposed to not-φ, it will rarely be the case that a moralist will be undecided about whether or not φ-ing is wrong in some of those circumstances but not others. So my account of indecision does not completely save the platitude that moral indecision about an action that is in the offing will lead to inaction etc., but does predict that it will hold to a large extent.

A moralist who expects to find herself in dispositionally specific circumstance C* in the near future, and who is norm-undecided between norms forbidding and norms requiring φ in C*, lacks a discretion-licencing preference to φ or not-φ in C*, and so would expect to violate trust qua moral entrusted party if she did so. So, as long as inaction was possible and by her lights permissible, we should expect her to delay action to minimize the chance of a violation. Indeterminacy in the expected weighted utilities (for a moralist, by that moralist’s lights) of conforming to different comprehensive norm-partitions can, I suppose, sometimes be resolved by thinking harder about the implications of conforming to them. Thus we should expect her to engage in moral reasoning in an effort to resolve her indecision.

A moralist who is undecided about whether φ-ing in C is permissible lacks a discretion-licencing preference to φ in C, and so would violate trust qua moral entrusted party if she did so. So, as long as inaction is possible and by her lights permissible, we should expect her to delay action to avoid such a violation. Indeterminacy in the expected weighted utilities of conforming to different comprehensive norm-partitions can, I suppose, sometimes be resolved by thinking harder about the implications of conforming to them. Thus we should expect her to engage in moral reasoning in an effort to resolve her indecision.

On my account, moral indecision necessarily involves norm-indecision, and norm-indecision depends on the state of one’s moral reasoning, or the reasoning whereby one arrives at voiceable preferences qua moralist. So my account saves the
third platitude about moral indecision, namely, that such indecision depends on the state of one’s moral reasoning. The fourth platitude concerns the relationship between indecision and confidence. I will say how my account of indecision saves this platitude when I have given my account of confidence. I turn to this latter task now.

Holding a basic moral judgement necessarily involves having a voiceable preference for conformity to some specific norm-partition. I will call a norm-partition that stands in this relation to a moral judgement J, N(J). I propose that a moralist’s degree of confidence in a basic moral judgement J is roughly the expected weighted utility for her, by her lights, of following N(J), whenever following N(J) is possible. (I shall sometimes omit this ‘whenever possible’ qualification in what follows.) That is to say, it is the expected weighted utility for her, by her lights, of her best preferred live atomic sub-option of following N(J), whenever following N(J) is possible. A live atomic sub-option of following N(J) is any option of following a comprehensive norm-partition entailing N(J). So a moralist’s confidence in J is roughly the expected weighted utility for her, by her lights, of following her best-preferred comprehensive norm-partition among those entailing N(J).

This is only rough, because a utility is not a good candidate to be a degree of confidence. To be analogous to degrees of belief in the way that appearances require, a degree of confidence in a moral judgement ought to be a point between a maximum and a minimum value.214 A judgement with the minimum value (to which, in deference to conventional treatments of degrees of belief, we may assign the number zero) is one in which a moralist couldn’t be any less confident, if she were rational. A judgement with the maximum value (conventionally, one) is one in which the moralist couldn’t be any more confident, if she were rational. I want to say that a moralist’s degree of confidence in J is greater the higher the expected weighted utility of following N(J) is for her by her lights, but I also want it to necessarily fall between (or on) these two points (again, if she is rational).

I tentatively propose the following. A moralist S has a confidence of zero in J just in case there is no comprehensive norm-partition M such that, if S were perfectly rational and logically omniscient, following M would have an expected weighted utility for S by S’s lights that was lower than the expected weighted utility that following N(J)

actually has for S by S’s lights, holding S’s basic desires fixed. S has a confidence of one in J just in case there is no comprehensive norm-partition M such that, if S were perfectly rational and logically omniscient, following M would have an expected weighted utility for S by S’s lights that was higher than the expected weighted utility that following N(J) actually has for S by S’s light, holding S’s basic desires fixed. These upper and lower bounds for S’s confidence levels would change if S’s basic altruistic desires, or the basic desires from which her simplicity desires were derived, changed.

Let us see how well this account of confidence in judgements of basic moral principle saves the platitudes about confidence that I listed earlier, insofar as they apply to judgements of basic principle. First, the account allows that a person can have degrees of confidence in a judgement that she doesn’t hold herself. This is because a person’s degree of confidence in moral judgement J is merely the normalized expected weighted utility for her by her lights in the norm-partition N(J) that one must hold if one is to accept J; the option of following a norm-partition will have an expected weighted utility for S by S’s lights even if that norm-partition is not one she holds—even if, that is, it is not entailed by her maximal decided norm-partition.

Second, the account predicts that a rational moralist will have a higher degree of confidence in a moral judgement she holds than in any inconsistent moral judgement. Suppose for reductio that a rational moralist S holds a moral judgment J, that there is some moral judgement ~J that is inconsistent with J, and that S is more confident in ~J than she is in J. This means that N(~J) has a higher expected weighted utility for S by S’s lights than N(J) does. This in turn means that S’s best-preferred comprehensive norm-partition out of those entailing N(~J) has a higher expected weighted utility for S by her lights than S’s best-preferred norm-partition out of those entailing N(J). For S to hold J is for her maximal decided norm-partition to entail N(J). A moralist’s maximal decided norm-partition is the most extensive norm-partition that is entailed by the conjunction of the comprehensive norm-partitions that she best prefers, and between which the expected weighted utilities for her by her lights fail to discriminate. So if S holds J, S’s best-preferred comprehensive norm-partition among those entailing N(J) must be entailed by the conjunction of the best preferred norms between which the expected weighted utilities for S by her lights fail to discriminate. So there cannot be any comprehensive norm-partition that S prefers to her best-preferred comprehensive norm-partition entailing N(J). Thus, contrary to our initial supposition, S’s best-preferred comprehensive norm-partition out of those entailing N(~J) cannot have a
higher expected weighted utility for S by her lights than her best preferred comprehensive norm-partition out of those entailing N(J).

Third, my account predicts that a moralist’s confidence in judgements of basic moral principle will depend on the state of her moral reasoning. This is obvious, since the expected weighted utilities for a moralist by her lights in different norm-partitions will depend on the state of her moral reasoning—her current grasp of the various more or less simple comprehensive norm-partitions that entail it, and of the prescriptions of those norm-partitions in various circumstances.

Finally, I shall address the platitude concerning the relationship between confidence and indecision that I omitted to address when I gave my account of the latter. The platitude is this: if a moralist is undecided between two inconsistent moral judgements about φ-ing in C, her confidence in one judgement will not be clearly higher or lower than her confidence in the other. My theory predicts only that this proposition will hold in most instances, not that it will always hold.

My account of indecision says that for a moralist S to be undecided between a judgement that φ-ing in (perhaps dispositionally non-specific) circumstance C is wrong and a judgement that φ-ing in C is permissible is for the expected weighted utilities for S by S’s lights to fail to discriminate between norms forbidding and norms requiring φ-ing in C when you are pre-trust disposed to φ in C. My account of confidence says, roughly, that a moralist S’s confidence in the judgement that φ-ing in C is wrong is the normalized expected weighted utility for S by her lights of norms forbidding φ-ing in C when you are pre-trust disposed to do so and φ-ing in C when you are not pre-trust disposed to do so, and that her confidence in the judgement that φ-ing in C is permissible is the normalized expected weighted utility for her by her lights of norms requiring φ-ing in C when you are pre-trust disposed to φ in C.

If it were not the case that judging that φ-ing in C is wrong involves having a voiceable preference for following norms forbidding φ-ing in C when you are not pre-trust disposed to do so—if it only involved preferring norms forbidding φ-ing in C when you are pre-trust disposed to do so, i.e., a preference that actually clashes with the preference involved in judging that φ-ing in C is permissible—my account would predict the platitude. This is because if you are undecided about φ-ing in C when you are pre-trust disposed to do so, then, necessarily, the expected weighted utility for you by your lights of conforming to norms requiring φ-ing in C when you are pre-trust disposed to do so will not be determinately higher or lower than that of conforming to
norms forbidding \( \varphi \)-ing in C when you are pre-trust disposed to do so. So neither of these options will clearly have a higher expected weighted utility for you by your lights. Since confidence is normalized expected weighted utility for a moralist by her lights, this means in turn that your confidence that \( \varphi \)-ing is permissible and will not clearly be higher than your confidence that \( \varphi \)-ing is wrong, nor vice versa.

Let us now drop our false supposition. Judging that \( \varphi \)-ing in C is wrong involves having a voiceable preference for following norms forbidding \( \varphi \)-ing in C when you are pre-trust disposed to do so, and when you aren’t. The option of following such norms cannot have a higher expected weighted utility for a rational moralist by her lights than the option of following norms forbidding \( \varphi \)-ing in C when you are pre-trust disposed to do so. This is because the former norms are ‘stronger’—which is to say, they are compatible with fewer, or ‘rule out’ more, comprehensive norm-partitions. More specifically, the former norms are \textit{strictly} stronger than the latter: the former rule out every comprehensive norm-partition that the latter rule out, plus some that the latter do not rule out. The expected weighted utility for a moralist by her lights of following a norm-partition is the expected weighted utility for her by her lights of following her best preferred comprehensive norm-partition compatible with that norm-partition. Thus if norm-partition N1 is strictly stronger than norm-partition N2, N1 cannot have a higher expected weighted utility for a moralist by her lights than N2, and may have a lower expected weighted utility for her by her lights, because it may rule out her best-preferred comprehensive norm-partition compatible with N1. So the norm-partition forbidding both \( \varphi \)-ing in C when you are pre-trust disposed to and \( \varphi \)-ing in C when you aren’t pre-trust disposed to may have a lower expected weighted utility for a moralist by her lights than the norm-partition merely forbidding \( \varphi \)-ing in C when you are pre-trust disposed to.

This means that, even though neither of the expected weighted utilities for a rational moralist by her lights of norms forbidding and norms requiring \( \varphi \)-ing in C when you are pre-trust disposed to will be clearly higher than the other if she is undecided about whether \( \varphi \)-ing in C is permissible, the expected weighted utility for her by her lights of norms forbidding both \( \varphi \)-ing in C when you are pre-trust disposed to and \( \varphi \)-ing in C when you aren’t pre-trust disposed to may be determinately lower than that of norms requiring \( \varphi \)-ing in C when you are pre-trust disposed to. Thus a moralist who is undecided about whether \( \varphi \)-ing in C is permissible may have more confidence in the
judgement that φ-ing in C is permissible than she has in the judgement that φ-ing in C is wrong.

This will occur when the moralist’s best-preferred sub-option of following the norm-partition forbidding φ-ing in C when you are pre-trust disposed to is the option of following a comprehensive norm-partition that both forbids φ-ing in C when you are pre-trust disposed to and requires φ-ing in C when you aren’t pre-trust disposed to (more carefully, it at least requires this on some precisification of C—I ignore this complication in what follows). For it is norms that make this latter requirement that are ruled out by norms forbidding both φ-ing in C when you are pre-trust disposed to and φ-ing in C when you aren’t pre-trust disposed to, but not by norms merely forbidding φ-ing in C when you are pre-trust disposed to.

It ought to be fairly unusual for a moralist to prefer norms both forbidding φ-ing in C when you are pre-trust disposed to and requiring φ-ing in C when you aren’t pre-trust disposed to to norms forbidding both φ-ing in C when you are pre-trust disposed to and φ-ing in C when you aren’t pre-trust disposed to. At least, this is the implication of the hypotheses about the nature of altruistic desires that I made in 4.1.4 with the aim of explaining the salience of permissibility judgements. So my theory does predict that a moralist who is undecided about the permissibility of φ-ing in C, and she finds the judgement φ-ing in C is required just in case you wouldn’t have φ-ed in C if left to your own devices more plausible than the judgement φ-ing in C is wrong, we don’t expect her to necessarily have the same confidence in the judgement that φ-ing in C is permissible as she has in the judgement that φ-ing C is wrong. Perhaps we would not be surprised in this case if she had a determinately lower confidence in the judgement that φ-ing in C is wrong than she had in the judgement that φ-ing in C is permissible. I find this sort of case quite difficult to grasp intuitively, and hence difficult to form clear intuitions about. I am inclined to doubt that those cases in which fiduciary non-cognitivism fails to save the ‘platitude’ amount to compelling counterexamples to the theory.
8.3 A Fiduciary Non-cognitivist Solution to the Puzzle

I will now turn my attention to the metaethical problem of moral uncertainty that I described at the start of this chapter. How do I explain the rationality of acting on moral judgements in which you are not completely certain? A simple and approximately correct answer to this question is as follows: the fiduciary non-cognitivist endorses the Instrumental Explanation. Moralists tend to act on their moral judgements, not because these somehow constitute their final ends, but for instrumental reasons. Acting on them ought to be a somewhat reliable way to satisfy their basic altruistic desires, but more importantly, it is by their lights the most likely means to satisfying their discretionary desires. When moralists rationally act on moral judgements that they hold with imperfect confidence, it is typically their discretionary desires that explain this. On my theory, moralists’ discretionary desires are, or are analogous to, desires to ‘do what is right’ in the de dicto sense.

Fiduciary non-cognitivism happily avoids the problem that I argued besets the version of the Instrumental Explanation that I discussed in Section 8.1. It does not predict that a woman who reaches the conclusion that abortion is permissible will nonetheless forbear from having one because she thinks that there is a tiny chance that she is wrong. Discretional desires do not prompt their possessors to maximize expected conformity to moral requirements. They prompt moralists to act on their discretion-licencing preferences—or, more accurately, the discretion-licencing preferences they would have if they were rational, though in practice this will usually lead to the same choices.

Rational moralists’ voiceable preferences depend on their degrees of belief as well as their desires—chiefly their credences in a priori propositions concerning the

215 Fiduciary non-cognitivism’s explanation doesn’t quite qualify as a version of the Instrumental Explanation, because it doesn’t predict that a rational moralist’s motivations will vary depending on her confidence in her moral judgements.

216 Michael Smith famously argued that the view that de dicto desires to do the right thing play an important role in moral motivation implies that (good) moralists make a fetish of acting morally, when intuitively they are reliably moved to do the right thing out of more intelligible motivations, such as (underived) desires to help people and relieve suffering (Smith, The Moral Problem, 75-6). This seems a good argument, not against the idea that a distinctively moral desire to do the right thing is explanatorily important, but against metaethical theories that fail to give an account of the content and provenance of this desire that makes it intelligible. It is not a good argument against fiduciary non-cognitivism, because discretionary desires do not seem fetishistic. A desire to live up to the trust of one’s fellow human beings is a perfectly intelligible one. (I have only addressed one facet of Smith’s fetishism argument here, the one that strikes me as the most initially convincing as an objection to fiduciary non-cognitivism.)
implications of following different comprehensive norm-partitions (e.g. beliefs about whether, in some particular cases, utilitarian norms require hanging the innocent). It might be thought that, if a moralist’s voiceable preferences hinged on very low credences in such propositions, and if a great frustration of altruistic desires were at present at stake, her discretionary desires might urge her to act on the norms that she would prefer if those propositions were false. This is not the case. A moralist’s discretionary desire will almost never prompt her to defy her norms. They urge her to choose the live atomic options with the unique highest expected weighted utility for her relative to her evidence. The live atomic options such that she believes they are most likely to have the unique highest expected weighted utility for her relative to her evidence, and those that she voiceably prefers, will almost always be the same. (If she is undecided about what to do in her present circumstances, she will lack a voiceable preference among her options, and she will probably also lack a belief, of any of the options among which she is undecided, that that option has the unique highest expected weighted utility for her relative to her admissible evidence. In such circumstances her discretionary desire will not prompt her to choose any particular option.)

According to fiduciary non-cognitivism, even though the woman in the abortion case judges that it is merely permissible to have an abortion, she would ‘in principle’ violate trust if she forbore from having an abortion given that she wanted to. This is because, if she judges that abortion is permissible, she has a voiceable preference to follow norms requiring her to have an abortion if she is pre-trust disposed to. So her discretionary desires will urge her to act on her pre-trust disposition. (In fact it is doubtful that she could violate trust by failing to act on her pre-trust disposition. This is because, given her voiceable preferences, no desire of hers to defy her pre-trust disposition could be a trust-based desire for her, at least qua moralist; and if it were a non-trust-based desire, and were strong enough to move her to action, it would determine the direction of her pre-trust disposition instead of opposing it. But what is important for my argument is that the distinctively moral motivation I have posited would not move a rational moralist to act on moral judgements other than those she holds, which is to say, it would not move her to act on norms that were not entailed by her maximal decided norm-partition.)

I have said that cognitivist defenders of the Instrumental Explanation can avoid making the wrong prediction in the abortion case by availing themselves of something like the non-cognitivist’s solution. Here’s what I have in mind. The cognitivist could,
like the fiduciary non-cognitivist, analyse permission judgements as requirement judgements of a special kind. These would be judgements that agents ought to do what they would want to do if moral propriety were not taken into account. On this view, to judge that abortion in circumstances C is permissible is to judge that an agent ought to have an abortion in C if she ‘pre-morally’ prefers to.\(^{217}\) The cognitivist could then posit that our \textit{de dicto} moral motivation is a standing desire to act in accordance with moral requirements, including the special requirements that permission judgements are concerned with. On this view, the \textit{de dicto} moral desire of the woman in the abortion case would presumably not move her to act on her tiny credence that abortion was wrong, because she would have a far greater credence that not having an abortion in her ‘conative circumstances’ was ‘wrong’, in a sense that was relevant to the satisfaction of the \textit{de dicto} desire.

\textit{Conclusion}

I have argued in this chapter that fiduciary non-cognitivism has an adequate account of moral uncertainty. Fiduciary non-cognitivism accommodates the most obvious facts about moral indecision and moral uncertainty. Moreover, it avoids an under-recognized problem facing cognitivists who hold that moral judgements are the result of reasoning about final ends, and that we commonly conform to them as such, without any ulterior motive.

\(^{217}\) If Michael Smith analysed permission judgements in this way, he could, I concede, avoid my criticism that he fails to explain our intuitions about the abortion case.
Moral Forgetting

Consider the following utterances:218

(1) Sorry I took your umbrella. I forgot that stealing was wrong.

(2) Is it good to give to international aid when your own family is in need? I’ve been racking my brains, but I can’t remember.

(3) You know that feeling you get when you can’t quite remember something that you desperately need to remember? That’s how I felt when Carla asked me whether it would be morally wrong for her to go through with the abortion. I knew the answer, but it had just slipped my mind! It was driving me crazy.

As these examples illustrate, claiming to have forgotten a moral truth can sound absurd. This demands an explanation. My chief concerns in this chapter will be to show that explaining our intuitions about utterances like (1)-(3), and related examples, is no easy feat, and that fiduciary non-cognitivism is among those theories that can well explain these intuitions. As we shall see, certain versions of moral non-cognitivism are peculiarly well placed to explain these phenomena. There are reasons to doubt that fiduciary non-cognitivism is among them, but I shall be argue that they are unconvincing. This chapter draws inspiration from observations and arguments made by Gilbert Ryle in his 1958 essay, ‘On Forgetting the Difference Between Right and Wrong.’219

218 A version of this chapter was previously published under the title ‘Forgetting Your Scruples’ in Philosophical Studies 173 (2016): 2889–2911.
219 The present paper aims to improve on Ryle’s argument chiefly in the following two respects. First, Ryle sets out to explain why it sounds absurd to say that one has forgotten ‘the difference between right and wrong’, which gives the impression that he is talking about a very general kind of moral forgetting, involving a wholesale loss or deterioration of moral knowledge. In fact the puzzle is far broader than this, for many quite specific reports of moral forgetting, such as (1)-(3), seem absurd. Ryle’s characterization of the problem thus makes it seem less troubling than it is. I address the broader puzzle. Second, I consider a wider range of rival explanations.
In what follows, whenever I talk of losing beliefs or other attitudes ‘through’ forgetting, I always mean that the loss is an instance of forgetting, not that it is a consequence of forgetting (something else). By ‘moral forgetting’ I always mean the loss of moral belief through forgetting. My use of ‘absurd’ is, I hope, close to the core folk usage. To be absurd, a thing must at least be odd; it must either resist intelligibility (for reasons other than complexity, indistinctness, or a lack of background information) or be ridiculous; it need not be conceptually impossible.

In the preceding chapters I have tried to be agnostic on the question of whether or not words like ‘true’, ‘believe’, ‘assert’, and ‘proposition’ should be given a deflationary construal, as many expressivists would have it. Thus I have avoided using these terms in connection with moral judgements and utterances. In this chapter it would be difficult to avoid using these folk terms. I shall be trying to elicit rather elusive pretheoretical intuitions about moral judgements, and this would be harder if I had a very limited ordinary-language vocabulary at my disposal. Moreover, for somewhat complex dialectical reasons, I face less pressure in what follows to limit my vocabulary in this way. So I shall no longer avoid using the ‘cognitivist-sounding’ terms in moral contexts.

In Section 9.1 I explore our intuitions about moral forgetting and defend a certain characterization of our explanandum. In 9.2, I evaluate several candidate explanations.

### 9.1 What Needs to Be Explained?

Our intuitions about (1)-(3) belong to a class of intuitions about moral forgetting for which a common explanation seems desirable. To explain these intuitions, we must try to identify this class. Our responses to (1)-(3) seem to justify the generalization that

---

Ryle does not (seriously) consider some cognitivist-friendly explanations for the phenomenon that present-day moral philosophers will find relatively credible, such as the argument from intuition (Section 9.2.3).

220 It seems that the versions of non-cognitivism that are best equipped to explain the phenomena are ones that embrace deflationism. If a version of non-cognitivism couldn’t explain talk of moral beliefs and truths, but correctly predicted that talk of moral forgetting would sound puzzling, this correct prediction might seem merely fortuitous rather than truly explanatory. So a policy of considering only those metaethical theories that seem best equipped to explain our intuitions about moral forgetting arguably permits me to consider only versions of non-cognitivism that are conjoined with deflationism about most ‘realist-’ or ‘cognitivist-sounding’ terms (or which can vindicate their use in moral contexts in other ways). If I consider only these theories, there is no need for me to avoid using the controversial terms.
reports of moral forgetting seem absurd. I’ll take this as a first-pass statement of our explanandum, and consider to what extent it must be revised.

Although I’ll be defending a particular statement of our explanandum in this section, it is not my plan to take this as fixed in Section 9.2, and rank candidate explanations according to their success at predicting it. The statement is revisable in principle. Most of the explanations I consider will make some predictions that are at odds with it, and I won’t count these predictions as defects without investigation.

9.1.1 Behavioural and Phenomenological Symptoms of Forgetting in Moral Contexts

For all my first-pass statement says, our explanandum might be merely linguistic, a puzzle about why sentences or speech acts reporting moral amnesia sound absurd. But I think it would be surprising if this were the case, and indeed there is evidently nothing particularly linguistic about our explanandum. Forgetting affects what we do and feel as well as what we say. And, just as reports of moral forgetting seem odd, so too do described cases in which a subject’s loss of moral belief is accompanied by the distinctive behavioural and phenomenological symptoms of forgetting.

Forgetting itself does not have a phenomenology, nor, typically, any immediate external symptoms. But efforts to remember and prevent forgetting are associated with distinctive behaviours and feelings. When we are worried that we are going to forget things, we repeat them to ourselves to commit them to memory, and test our recollection of them. We leave reminders for ourselves. We tie strings around our fingers. When something we urgently need to remember has slipped our minds we sometimes have a frustrated ‘tip-of-one’s-tongue’ feeling, we ‘rack our brains’, we express and exhibit frustration. When we recall a forgotten truth, we have a feeling of having recovered something lost, which is quite different from (for instance) any feelings associated with re-adopting a belief that one had earlier rejected.

Imagining cases in which these behaviours and feelings are associated with efforts to maintain or re-form moral beliefs produces a sense of incongruity. Moral reminders seem odd—it would be bizarre if Carla left a reminder on her fridge door that abortion is permissible, reasoning that this is information she would like to have at hand if she decided she wanted to terminate her pregnancy. It would be bizarre to repeat moral truths to oneself in order to memorize them, and quiz oneself to test one’s recollection. And the behaviours and feelings reported by the speakers of (2) and (3) in connection with their efforts to re-form their lost moral beliefs seem similarly bizarre.
Apparently, all indicators of actual or potential moral forgetting (as opposed to mere moral belief-loss) are apt to puzzle us. Our explanandum is thus not merely linguistic. It seems that we want to explain why *moral forgetting seems absurd*, not just why reporting moral forgetting seems absurd.

9.1.2 *Forgetting Morally Relevant Non-normative Facts*

However, there appears to be a large and important class of counterexamples to this characterization of our explanandum. The following utterances do not sound absurd:

(4) I forgot that John was such a bad man.

(5) I should thank Ellie again for helping me get this job. I keep forgetting what a good deed that was.

(6) Yes, you told me what Jessica did, but that was three days ago. I can’t remember whether or not what she did was immoral.

What distinguishes these sensible-sounding utterances from puzzling reports of moral forgetting like (1)-(3)? It seems likely that the reason why (4)-(6) sound sensible has something to do with what these utterances communicate to us about the sorts of facts the speakers have forgotten.

We naturally infer from (4)-(6) that the speakers have forgotten what they take to be morally important *non-normative* facts about the things whose moral status they claim to have forgotten. For instance, it might be natural to suppose that the speaker of (4) has forgotten about some of John’s behavioural dispositions, that the speaker of (5) keeps forgetting the inconvenience that Ellie had to put up with in helping the speaker get her job, and that the speaker of (6) has simply forgotten what Jessica did. By contrast, we are not inclined to infer that the speakers of (1)-(3) have forgotten morally important non-normative facts about stealing, giving to charity when one’s family is in need, or abortion, respectively.

Suppose that a person believes both that some object *x* has some non-normative property *F*, and that *x*’s being *F* makes *x* have some moral property *M*. Presumably, if she forgot that *x* is *F*, then she might cease to believe that *x* is *M* as a result. She might rationally reject this moral belief, or the moral belief, having lost the rational support which would have kept it in existence, might just cease through some non-rational process. Be it rational or non-rational, I shall call this kind of cessation of moral belief *Loss of moral belief Resulting from Non-normative Forgetting* (LRNF).
We should not expect LRNF to seem absurd, and I suggest that (4)-(6) do not seem absurd because, unlike (1)-(3), we take them merely to be reports of LRNF. I hypothesize that some utterances which have the appearance of reports of moral forgetting are actually reports of LRNF; that is, some sorts of utterance which have the form of reports of moral forgetting are characteristically used to report LRNF, and are interpreted as reports of LRNF by competent language-users. Utterances (4)-(6) are utterances of this sort. I hypothesize that there are certain properties (certain types of content, linguistic cues, and the like) that an apparent report of moral forgetting may possess which conventionally indicate that it is a report of LRNF, and that (4)-(6) have these properties.

Though I’ve called such utterances ‘apparent’ reports of moral forgetting, I don’t wish to deny that they may actually be reports of moral forgetting. But if they are genuine reports of moral forgetting, and do not seem absurd, it is because LRNF sometimes counts as moral forgetting—that is, losing a moral belief that \( p \) as a result of forgetting what one takes to be a morally relevant non-normative fact can count as forgetting that \( p \). Whether or not it can is a conceptual question that I shall not address.

We can now refine our explanandum. We do not want to explain why moral forgetting seems absurd, because this may not be true. We want to explain why moral forgetting that is not LRNF seems absurd.

9.1.3 Inverted-Commas Moral Forgetting

Familiarly, there is a distinction to be made between inverted-commas moral beliefs and non-inverted-commas (or ‘genuine’) moral beliefs. The former are psychological beliefs about the moral beliefs of other people. They are so called because they can properly be expressed by ordinary moral sentences with the moral terms in inverted commas. The inverted commas need not be explicit (or even ‘conscious’). For instance, ‘Ashley won’t come to the party because she’s too virtuous’ may express, among other things, the belief that Ashley would be considered virtuous to a certain degree by some salient persons.\(^{221}\)

Reports of inverted-commas moral forgetting do not generally sound absurd. We would not be puzzled if an anthropologist, discussing the moral code of a foreign culture with another anthropologist, were to ask, ‘Is it wrong to marry any of your

---

\(^{221}\) For influential discussions, see Hare, *Language of Morals*, 124-6, and 167 (on the ‘unconscious’ inverted-commas use of value terms).
uncles, or just consanguine uncles? I’ve forgotten.’ Does this mean I should revise my statement of our explanandum to exclude inverted-commas moral forgetting? This example does not give me much reason to. There is little temptation to regard the anthropologist’s utterance as a report of moral forgetting, because we are not inclined to think that her lost belief was a moral belief.

Some inverted-commas cases are more ambiguous, however. Suppose that an old-fashioned school teacher has taken it upon herself to provide her young pupils with moral instruction. Her method is simply to testify that a particular set of moral principles is true. The children take her moral principles seriously; they denounce other children who violate them, and feel guilty when they themselves violate them. Observe that we can, without puzzlement, imagine the children memorizing, remembering, and forgetting the teacher’s moral principles. We can, without puzzlement, imagine the teacher quizzing a child on the content of a previous lesson—‘Is it morally wrong to covet?’—and the frustrated child replying, ‘I can’t remember!’

It is clearly compatible with the description of this case that the only beliefs the teacher ever succeeds in instilling in her students through her moral instruction are inverted-commas moral beliefs—e.g. beliefs about the moral opinions of the teacher. Their apparently moral emotions and motivations might derive in more or less complex ways from their desire for the teacher’s approval. Indeed not only is an inverted-commas interpretation of the case available, it also seems very natural, and I submit that this explains why the case is not puzzling. But even if it were stipulated that the instilled beliefs were inverted-commas moral beliefs, it would not obviously be wrong to think that they were moral beliefs on a folk understanding. They bear considerable resemblance to paradigm instances of moral belief, and the resemblance may be sufficiently close that they fall under the folk concept of moral belief. If so, the case is a counterexample to the claim that moral forgetting seems absurd. So I shall revise my statement of our explanandum to exclude inverted-commas moral forgetting.

Here is our new statement in full: moral forgetting seems absurd, setting aside LRNF and inverted-commas moral forgetting. Since I’ll have little more to say about LRNF and inverted-commas cases, I won’t generally bother to include this qualifying clause when I state our explanandum. Also, when I talk about moral belief it can be assumed that I have in mind only non-inverted-commas moral belief, and that when I talk about moral forgetting I have in mind only moral forgetting that is not merely LRNF, or inverted-commas moral forgetting.
This is my final statement of our explanandum. I’ll now see how well it fares against a few challenges.

9.1.4. *Rhetorical Reports of Moral Forgetting*

It might be thought that the naturalness of utterances like the following shows that it is not true in general that moral forgetting seems absurd:

(7) Through years of exposure to corporate pressures and the incessant call of the rat race, he simply forgot that other people count.

However, I think that (7) sounds natural only because we don’t take it literally. Ironic and metaphorical reports of moral forgetting have some currency in polemical, moralistic discourse. Some examples from the Internet:

Mr Ammiano in his concern about trampling the rights of the LGBT community seems to have forgotten that the OTHER students have rights as well. Don’t the girls in a locker room have the right to NOT be exposed in front of members of the opposite sex?²²²

If we had not forgotten that it is wrong to steal, we would not be faced today with a kleptocratic government ….²²³

In such reports, the speaker typically presents moral forgetting as an explanation for actions, attitudes, or utterances which she opposes or finds regrettable for the very reason that they violate or contradict the supposedly forgotten moral requirements. Thus rhetorical reports of moral forgetting rarely attribute amnesia to the speaker herself at the time of utterance. If (7) seems natural because the story it tells seems natural when taken literally, and not because we interpret (7) as rhetorical, then this story should not seem absurd when we retell it (or imagine it) from the point of view of the forgetter before he has recovered from his amnesia, as in the following:

(8) Through my years of exposure to corporate pressures and the incessant call of the rat race, I’ve simply forgotten whether or not other people count.

This example lacks the polemical overtones of (7), and it sounds much less natural than (7). Moreover, if we imagine the speaker racking his brains and experiencing tip-of-one’s-tongue feelings as he tries to recall whether or not other people count, the

---


described forgetting seems quite absurd. Thus (7) doesn’t seem to be a counterexample to our explanandum.

As the contrast between (7) and (8) suggests, among reports of moral forgetting, we are more than usually willing to interpret third-person reports as rhetorical. It is mainly for this reason that all of the reports of moral forgetting that I cite in support of our explanandum elsewhere in this chapter are first-personal. This does not mean that third-person reports are not puzzling, only that more care must be taken when devising third-person examples to discourage non-literal interpretations. Such discouragement may come in several forms. We are less likely to interpret a third-person report of moral forgetting as rhetorical if the reporter herself admits, especially at the time of utterance, to being ignorant of the forgotten truth; if the forgotten truth is unobvious; if the tone of the report is dispassionate; if the forgetter’s doubt is emphasized; or if the report somehow contrasts the forgetting with the sort of blitheness or self-deception that is called forgetting in rhetorical reports.

Here are some absurd-seeming third-person reports of moral forgetting that take advantage of these facts to varying degrees.

(9) Things would certainly be easier if we ended the pregnancy. But I’m not sure whether abortion is wrong, and Carla can’t remember.

(10) I told him to think about what he was doing, but he didn’t want to hear. It saddened me. The Eddie I knew would never have had an affair with a married woman. He always knew that it was wrong. Well, except for that time with Laura, but on that occasion I think he’d just forgotten. Eddie’s pretty forgetful.

(11) Jane’s husband repeatedly urged her to protest against the death penalty in her speech at the Civil Rights Society luncheon. But Jane prevaricated, and gave excuses for avoiding the issue. The truth was that she doubted that her husband really knew whether capital punishment was wrong, and she didn’t know herself because she couldn’t remember.

9.1.5 Moral Forgetting Caused by the Forgetting of Moral Arguments

One reason why the seeming absurdity of moral forgetting is surprising, and why it stands in particular need of explanation, is that our moral beliefs often have a theoretical character, and are the result of careful thought and argument. We might expect to be puzzled if a person lost through forgetting a belief that was directly supported by her present observations, sensations, or intuitions. But theoretical beliefs do not generally have this sort of direct support, so we should not expect their loss through forgetting to puzzle us.
Perhaps, though, this characteristic of moral beliefs makes the generalization that moral forgetting seems absurd a bit too surprising, inviting scepticism. Investigation may reveal that reporting moral forgetting does not seem absurd if the context informs us that the lost moral belief was theoretical rather than directly intuitive.

This idea will reappear in 9.2, when I consider a couple of explanations for our intuitions about moral forgetting that do not predict that cases involving the loss of theoretical moral beliefs through forgetting will seem absurd (or so I argue). At this point I want to consider a particular version of the idea which seems to me relatively promising as a challenge to our explanandum. Observe that there doesn’t seem to be anything puzzling about forgetting arguments for moral claims. It might therefore seem unlikely that we would be puzzled by cases in which a person forgets a moral truth because she forgets an argument establishing it. If so, our explanandum must be further qualified.

Admittedly, some apparently straightforward tests of the hypothesis that moral forgetting resulting from the forgetting of moral arguments is not puzzling seem to support that hypothesis. For instance:

(12) I forgot that abortion was wrong because I forgot the argument on the basis of which I believed that abortion was wrong. I don’t have a strong intuition of absurdity when I read (12), but I also feel that I have only a superficial grasp of its meaning. The sentence hardly makes the reported psychological events vivid. We would do better to consider a more colourful example.

The following example first illustrates what it means to forget a philosophical argument on the basis of which one holds a moral belief, and then illustrates how things might be if, having forgotten the argument, one were to lose the belief through forgetting.

(13) Don Marquis has an argument which shows that killing people is wrong because it deprives them of their future experiences and projects. Hence killing unborn people is also wrong.\footnote{Cf. Don Marquis, ‘Why Abortion is Immoral’, \textit{Journal of Philosophy} 86 (1989): 183-202, 189-92.} This is the argument that made me realize, years ago, that abortion is wrong. Unfortunately I don’t have a good memory for philosophical arguments, and I’ve been embarrassed more than once in debates with defenders of abortion by my inability to remember why we should believe that abortion is wrong. Things recently reached a new low, however. Some acquaintances of mine were arguing about the ethics of abortion. I was called upon to give my opinion on the subject, but I found, to my frustration, that I didn’t know what to tell them. This time, not
only I had I forgotten Marquis’s argument, but I’d forgotten that abortion was wrong too.

This report of moral forgetting seems absurd, despite the speaker’s admission that she had forgotten the argument supporting her lost moral belief. So I don’t see any reason to weaken our explanandum by excluding cases in which the forgetting of moral arguments is available, or made salient, as a potential explanation for moral forgetting.

9.1.6 Complexity

Complex truths are generally easy to forget, so it might seem unlikely that the forgetting of complex moral truths would seem absurd. When testing this idea, we must be wary of using examples in which someone is unable to remember whether or not some complex moral principle is true, where the principle is picked out by its name or some identifying feature other than a full statement of its content. For we may be tempted to think the subject has merely forgotten the truth of the principle as a result of forgetting what the principle says, and this would be an instance of LRNF. Similarly, we must avoid examples in which someone has forgotten the moral status of an action in some complicated case, where this would naturally be explained by her inability to recall details of the case itself.

I think the following example avoids these problems:

(14) Mark wanted to know about the Doctrine of Triple Effect. I explained it to him in the following way. Suppose that (a) an agent A expects that an action $\phi$ would have both a good outcome and a lesser evil outcome; (b) this lesser evil outcome is such that it would be wrong to intend it; (c) A $\phi$-s, taking both the expected good outcome and the expected lesser evil outcome to be conditions of her action, but not intending either; and (d) A’s $\phi$-ing does in fact cause the good outcome. According to the Doctrine of Triple Effect, A’s $\phi$-ing is not necessarily wrong, even though A does not intend the good, and would not have $\phi$-ed had she not expected to bring about the evil. Mark found this very interesting, and wondered about the truth of the doctrine. Was A’s action necessarily wrong? Mark was inclined to think that it wasn’t. As for me, I was at a loss. I just couldn’t remember.

Given her confident grasp of the case in question, and despite its complexity, the speaker’s moral forgetting seems very odd. This is consistent with my current formulation of our explanandum.


226 The naturalness of the following seems to lend credence to the complexity worry: ‘I believe that the Doctrine of Triple Effect is true, but it’s a complex doctrine and I can’t remember all the details of it.’ This is not a clear counterexample to our explanandum, however. The speaker
9.1.7 Memory Disorders

It would be very surprising if the generalization that moral forgetting seems absurd held even for cases involving people with memory disorders, such as Alzheimer’s sufferers and stroke victims. And yet it may hold even in these cases. Consider the following exchange between an old man and a care worker.

A What’s all this rubbish doing here?
B It’s from your daughter’s visit this morning.
A I don’t remember that.
B It’s your birthday. Your daughter and her family came, and you all sang and ate cake.
A My birthday? How old am I then?
B Eighty-seven.
A Good Lord! Eighty-seven! What did Kate give me?
B She gave you this card, and a hat, and a calendar. Look.
A [Examines the gifts.] What rubbish!
B That’s not what you said when she gave them to you. You were all smiles and thank-yous. Then when she left you threw them away.
A Well, I suppose I was being polite.
B Polite or not, you shouldn’t deceive people.
A Most people think it’s alright to lie to spare someone’s feelings.
B But are they right?
A Hmm. Actually, I can’t remember.

A evidently has a memory disorder, yet his report of moral forgetting seems puzzling. So I am not inclined to view memory disorder cases as counterexamples to our explanandum.

On the other hand, I had much more trouble than usual devising a memory disorder case that elicits the typical intuition. There are many that do not, and while I suspect that this can be explained away, I lack the space here to establish this. I thus take the matter of whether memory disorder cases support our explanandum to be in doubt (more so than usual). When considering explanations for the puzzle of moral forgetting in the next section, I won’t count a ‘failure’ to predict that memory-disorder cases are puzzling as a defect of a candidate explanation.

may have a moral belief whose normative content is exactly \( p \), where \( p \)’s resemblance to the Doctrine of Triple Effect is imperfect but close enough that she counts as believing the Doctrine in virtue of having this moral belief. This being so, she may have forgotten some details about what the Doctrine says without having lost any moral beliefs through forgetting. Thus we can interpret her as speaking truly and literally without supposing that she has experienced (non-LRNF) moral forgetting.
9.2 Explanations

I’ll now evaluate some explanations for the puzzle of moral forgetting. The explanations are roughly organized into two groups: those in subsections 9.2.1-4 emphasize epistemic considerations; those in 9.2.5-7 emphasize connections between our moral beliefs and our cares and concerns. Most of the explanations assume a commonsense cognitivist metaethic; exceptions are indicated.

9.2.1 Commonsense Moral Scepticism and the Factivity of ‘Forget’

Though not plausible, the following explanation for the apparent absurdity of reports of moral forgetting is worth considering because it is bound to suggest itself to anyone who gives any thought to the issue, and can be something of a distraction. ‘Forget’ is factive: one cannot forget or remember that \( p \) unless \( p \). Moreover, to forget is plausibly to lose knowledge, not merely true belief: one cannot forget that \( p \) unless one knows that \( p \). This being so, it might be thought that the reason why reports of moral forgetting sound odd is that they imply that the forgetter possessed moral knowledge. This implication seems absurd (we are to suppose) because the idea that anyone actually possesses moral knowledge defies common sense. The phenomenon of widespread and intractable moral disagreement is clear and compelling evidence that none of our moral beliefs is true save by accident. Our moral beliefs are merely ‘personal opinions’, not knowledge, and, though we may sometimes pretend otherwise, we all basically understand this.\(^{227}\)

The most obvious problem with the commonsense moral scepticism explanation is that moral knowledge does not seem absurd. Compare:

(15) I knew that abortion was wrong.
(16) I knew that abortion was wrong, but then I forgot.

(15) sounds fine; (16) sounds absurd. The hypothesis that we are all moral sceptics deep down fails to explain our differing responses to these examples.

\(^{227}\) This hypothesis is only a candidate for an explanation of our intuitions about reports of moral forgetting. But it might also point toward a certain sort of explanation for the non-linguistic phenomena. Why are we puzzled by cases in which people rack their brains in order to recover lost moral beliefs, or try to preserve their moral beliefs by memorizing or leaving reminders? It might be said that this is because common sense tells us that we have no reason to think our moral beliefs true, and hence no reason to regret or prevent their loss.
9.2.2 Too Obvious to Forget

It’s not surprising that (1) strikes us as ridiculous. How could anyone forget something as obvious as the fact that it’s wrong to steal? Perhaps it is the obviousness of moral truths that explains why it seems absurd to say that one has forgotten them.

This is another thought that a casual reflection on our problem might give rise to, but which clearly cannot be right. It isn’t obvious whether or not abortion, capital punishment, euthanasia, or killing animals for food are wrong. Yet (16) seems absurd, and retains this appearance when we substitute ‘capital punishment’ etc. for ‘abortion’. Anyway, if reports of moral forgetting sound absurd because the requirements of morality are obvious, I cannot see why reports of uncertainty regarding these requirements should not seem absurd. Yet there does not seem to be anything absurd about being uncertain whether or not even stealing is wrong.

9.2.3 Intuition

Consider the following argument:

P1 We assume that all basic moral beliefs are supported purely by intuition.

P2 The loss through forgetting of any belief that we assume is supported purely by intuition seems absurd.

C So the loss of all basic moral beliefs through forgetting seems absurd.

By ‘basic moral beliefs’ I mean moral beliefs that do not depend for their rationality or justification on beliefs about contingent matters. The phrase ‘supported purely by intuition’ in P1 and P2 is not meant to rule out a role for reasoning in the formation of the beliefs in question. I’m assuming that the role of theoretical reasoning is not to support beliefs, but to identify relationships of support between considerations and beliefs.

Observe that it’s not immediately clear whether C predicts, as our explanandum does, that all moral forgetting that is not LRNF or inverted-commas moral forgetting will seem absurd. However, setting aside inverted-commas cases, it is plausible that the vast majority of (apparent) reports of moral forgetting that do not seem to imply the loss through forgetting of basic moral belief will be those we interpret as reports of LRNF, which do not seem absurd anyway. So C may well be general enough to explain the phenomena. At any rate, I shall confine my attention to determining the adequacy, not of the generality of C, but of the argument for C.
Why accept P1? Most of the time at least, we arrive at our moral views by consulting our intuitions. So perhaps we assume that people’s moral beliefs are based on intuition because it is generally safe to do so. Why accept P2? There is some plausibility in the idea that beliefs that are supported by intuition cannot (normally) be lost through forgetting because they are being continually ‘propped up’ by those intuitions. Since both P1 and P2 have some credibility, we have here a candidate for a cognitivist-friendly explanation of our intuitions about reports of moral forgetting.

P1 is not obviously true. Don’t we sometimes form basic moral beliefs on the basis of testimony? If so, why do we assume that basic moral beliefs are supported by intuition? And even if we do typically assume this, why don’t we drop the assumption when we are presented with a case of moral forgetting, rather than interpret it in a way that seems absurd?

P2 appears to be false. Plausibly, if P2 is true, then the loss through forgetting of any belief that we know (rather than assume) is supported purely by intuition also seems absurd. But some beliefs that we know to have been formed purely on the basis of intuition can apparently be lost through forgetting, without this seeming absurd. I have in mind beliefs in propositions that are not themselves intuitively plausible, but which we hold because they are entailed by other propositions that are intuitively plausible. Consider the case of Eric, a native English-speaker who is persuaded that sentences like ‘Jack went to the party with Jill and I’ are ungrammatical by the following intuitive argument:

P1* ‘Jack went to the party with I’ is ungrammatical.

P2* If ‘Jack went to the party with I’ is ungrammatical, then sentences like ‘Jack went to the party with Jill and I’ are ungrammatical.

C* Sentences like ‘Jack went to the party with Jill and I’ are ungrammatical.

P1* and P2* are, I trust, intuitively plausible to native speakers. Eric himself accepts them purely on the basis of intuition, and accepts C* purely on the basis of a deduction from them. Years pass, and Eric is heard to say,

(17) Is it “Jack went to the party with Jill and I,” or “Jack went to the party with Jill and me”? I’ve forgotten.

228 It would be quite unusual for a cognitivist to deny this. A non-cognitivist might be tempted to, however. This is because, firstly, it is hard for a non-cognitivist to explain how moral truths could be communicated from one person to another in a rational way. It is not obvious that there is any rational process by which we can communicate desire-like attitudes to one another. Second, it’s not clear that denying that we form moral beliefs on the basis of testimony has a significant cost (cf. McGrath, ‘Skepticism About Moral Expertise’, 2011).
Eric’s report does not seem odd, even when we bear in mind that his reportedly forgotten belief was supported purely by intuition.

Cases like Eric’s are germane because many of our moral beliefs have this sort of derived intuitive justification. Consider the case of Derek, who is persuaded that the distinction between killing and letting die has no moral relevance when he is presented with the following pair of examples from James Rachels:

a woman [A] wants her uncle dead, and she gives him poison in his coffee. Another woman [B], who also wants her uncle dead, is about to give him poison when she sees him unknowingly drink poison from another source. She watches him die, withholding the antidote in her pocket.229

Derek reasons as follows:

P1′ A’s conduct is no worse than B’s.
P2′ If A’s conduct is no worse than B’s, then killing is no worse than letting die.
C′ Killing is no worse than letting die.
Derek accepts P1′ and P2′ purely on the basis of intuition, and accepts C* purely on the basis of a deduction from these premises. Years pass, and Derek is heard to say,

(18) Is killing worse than letting die? I’ve forgotten.

This report of moral forgetting would seem odd in any non-inverted-commas context, and it seems odd in Derek’s. Like C*, C′ is not, considered by itself, very intuitively plausible, though it is implied by intuitively plausible premises. But it is still puzzling when someone who believes C′ on the basis of an inference from these intuitive premises reports that he or she has forgotten C′. Perhaps Derek could have non-rationally ceased to believe C′ as a result of forgetting his argument for C′, but there seems to be something wrong with describing such a cessation of moral belief as forgetting.

The argument from P1 and P2 to C is thus unsatisfactory, and I can’t think of an improved version of this argument that would avoid the problems I’ve raised.

9.2.4 Speaker Subjectivism

According to speaker subjectivism, a (simple, indicative) moral sentence is a report concerning the psychology of the speaker. To illustrate, let me introduce T, a toy

Speaker subjectivism theory. T explains the meaning of (19) by postulating that it is identical to that of (20):

(19) Abortion is wrong.

(20) I disapprove of abortion.

Speaker subjectivism is normally taken to be a purely semantic thesis. But there is an obvious way in which a speaker subjectivist might complement her semantics with a moral psychology: she might hold that moral beliefs are the mental states expressed by moral sentences. For example, an adherent of T might hold that for a person to believe (19) is just for her to believe that she disapproves of abortion.

If speaker subjectivism is true, and if the mental states that are reported by moral sentences can easily be discovered through introspection, it looks as though we can explain why moral forgetting seems absurd. T explains the meaning of (21) by holding that it is identical to that of (22):

(21) I’ve forgotten whether or not abortion is wrong.

(22) I’ve forgotten whether or not I disapprove of abortion.

Thus if T is true, (21) presumably sounds ridiculous for the same reason that (22) does. That is, because it is generally harder to forget easily introspectable facts about one’s own psychology than it is to forget facts about the external world, and (less importantly) because there is not usually any point in telling people that one has forgotten a fact that one could rediscover through a moment’s introspection.

A proponent of T who coupled her semantics with a moral psychology in the manner described above could provide a parallel explanation for the seeming absurdity of cases in which distinctively mnemonic behaviours and feelings follow from a loss of moral belief. She need merely point out that to apparently lose a belief about what one disapproves of through forgetting seems bizarre even when this appearance is generated solely by a description of the behaviours and feelings caused by the belief-loss.

Unfortunately, the speaker subjectivist hypothesis is not a satisfactory and adequately general explanation for the phenomena. T holds that (23) has the same meaning as (24):

(23) I spent days trying to decide whether or not abortion is wrong.
(24) I spent days trying to decide whether or not I disapproved of abortion.²³⁰

In ordinary language, (24) can be interpreted in two ways, one literal, the other non-literal (but more natural). I’ll stipulate that T gives the literal interpretation of (24) as its translation of (23). On this reading (24) is a report that the speaker spent days engaging in psychological inquiry in order to find out whether or not she disapproved of abortion. Facts about your psychology that you can only discover through days of inquiry are facts that you can easily forget—at least, there is in general no unusual obstacle to your forgetting them. Such facts will not be easily introspectable; they will be more akin to the facts a professional psychologist would postulate to explain your thoughts, feelings, and behaviour.

It is thus doubtful that speaker subjectivism can offer a good explanation for the absurdity of many reports of moral forgetting, such as:

(25) I spent days trying to decide whether or not abortion is wrong. But when I finally knew the answer, I was summoned to sit on a jury, and had to turn my attention to other moral problems. By the time the trial was over, I really needed to decide whether or not to terminate my pregnancy, but I found I couldn’t remember whether or not abortion was wrong.

T would translate (25) thus:

(26) I spent days trying to decide whether or not I disapproved of abortion. But when I finally knew the answer, I was summoned to sit on a jury, and had to turn my attention to other moral problems. By the time the trial was over, I really needed to decide whether or not to terminate my pregnancy, but I found I couldn’t remember whether or not I disapproved of abortion.

Admittedly, (26) does sound very odd, for a variety of reasons. But if T is true, and if we bear in mind that the speaker’s belief about her attitude to abortion is theoretical rather than directly introspective—as we must do if the first sentence of (26) is to make sense—then there shouldn’t be anything to be puzzled about in the speaker’s admission that she lost this belief through forgetting.

Now, as a matter of fact there is something puzzling—a residual oddness—about this admission even when we bear in mind the theoretical nature of the belief, just as there is a residual oddness about the utterance, ‘I forgot that I hate my stepfather’, even when we know that the speaker only ever believed that he hated his stepfather on the basis of his psychologist’s testimony. But this is surely due to the fact that it is quite

²³⁰ This translation may be inadequate given the past tense; (24) should perhaps read, ‘I spent days deliberating on the question of whether or not I (then) disapproved of (and now disapprove of) abortion.’ Alterations of this sort will not importantly affect my argument.
unusual for people’s beliefs about their affective states to have this sort of inferential justification. If T were true, however, having an inference-based belief about what one approves of ought to be a very ordinary condition, no less ordinary than that of holding a moral belief that one does not find obviously true. So forgetting what one approves of presumably ought to be a common and familiar occurrence, and not a puzzling one. An adherent of T cannot point to the residual oddness of the report of psychological forgetting in (26) to explain the puzzling character of (25) unless she can explain why this residual oddness is present.

Recall that (24), in ordinary speech, has an alternative interpretation. If someone utters (24) she normally means that she spent days engaging, not in psychological inquiry, but in a kind of mental activity that commonly has as one of its outcomes the formation in the subject of a new attitude of approval, disapproval, or tolerance toward something. This mental activity is ordinarily called ‘moral reasoning’, but I shall call it approval-pumping. As a piece of ordinary speech, (24) would normally be taken to mean, in my terminology,

(27) I spent days engaging in approval-pumping vis-à-vis abortion.

It might be wondered whether a speaker subjectivist could plausibly give (27) instead of (24) as her translation of (23). To know this, we must have some idea what approval-pumping is for a speaker subjectivist. I assume that a speaker subjectivist would favour a characterization of approval-pumping similar to that which a pure expressivist or an emotivist might offer. That is, she would view it as a fundamentally non-cognitive mental activity. She would not hold, as other cognitivists might, that approval-pumping is a kind of doxastic reasoning, and that the changes in a person’s approval states that it leads to are the result of changes in her moral beliefs. A speaker subjectivist is in no position to claim that our moral beliefs alter what we disapprove of, when she also claims that those beliefs are about what we disapprove of.

This being so, the speaker subjectivist who gave (27) as her translation of (23) would incur a high cost. Observe that replacing the first sentence of (25) or (26) with (27) yields,

(28) I spent days engaging in approval-pumping vis-à-vis abortion. But when I finally knew the answer….

231 ‘Moral reasoning’ suggests much about the nature of this approval-affecting activity which is not metaethically neutral, and which might be misleading in the present context (e.g. it suggests that the activity is a kind of reasoning).
‘Knew the answer’? What could this mean? If speaker subjectivism is true, approval-pumping is not a search for answers (answers to what?), and its output is just an approval-type attitude, not knowledge. The corresponding parts of (25) and (26) make good sense, but (28) is baffling. Such implications give us reason to reject this version of speaker subjectivism. It looks as though speaker subjectivism is incapable of providing a general explanation for the puzzle of moral forgetting unless it is modified in ways that greatly diminish its plausibility.

9.2.5 Too Important to Forget

Moral facts often matter a great deal to people, and it can be surprising when someone forgets a fact that matters to her. Perhaps this explains why moral forgetting is puzzling.

This explanation obviously needs refinement. People forget things that matter to them all the time—this is why forgetting is often frustrating. It’s not normally a puzzling phenomenon. Moreover, part of the reason why forgetting facts that matter to us is comparatively unusual, and hence surprising to the extent that it is, is that we often make a greater conscious effort to remember such facts—for instance, by mentally reminding ourselves of them, or leaving external reminders, or memorizing. Yet this sort of conscious effort would itself seem bizarre in a case involving moral belief.

However, there are some important non-normative truths whose forgetting is quite robustly puzzling. The following would sound very strange in almost any context:

(29) I forgot that God exists.

Remembering whether God exists also rarely seems to require or involve effortful self-reminding or memorizing. And it is plausible that this is because it matters a great deal to people whether or not God exists. (30) might seem to provide proponents of the ‘too important’ explanation with grounds for thinking that a successful refinement of that explanation is waiting to be discovered.

Unfortunately, a crucial premise of the ‘too important’ explanation is false. It is not plausible that, in every puzzling case of moral forgetting, the forgotten moral fact mattered greatly to the forgetter (or that we assume that it did). A person may have an opinion about whether or not some moral claim \( p \) is true, and yet not care much whether or not \( p \). This is not a puzzling state of affairs. Yet we would still find it odd if she subsequently claimed to have forgotten whether or not \( p \). Consider:
I let Carla know that the abortion question had never mattered to or interested me, but she still insisted on hearing my opinion on it. Unfortunately I couldn’t tell her whether or not abortion is wrong, because I’d forgotten.

The first sentence of (31) is not particularly puzzling, but the report of moral forgetting sounds absurd in the familiar way.

Perhaps we doubt that the speaker of (31) is sincere when she claims not to care about the ethics of abortion, so let me make the same point again, appealing to different evidence. It is surely not true that almost every moral question matters to almost everyone. There are countless obscure moral questions that few people have thought important enough to have formed opinions on. Yet it still sounds odd when someone reports having forgotten the answer to such a question. Consider:

(31)  Tim and I both agreed that it’s permissible to kill in self-defence even when one’s attacker is an innocent person acting under hypnosis. But when Tim asked me if it would be permissible for a murderer to kill an innocent, hypnotized aggressor in self-defence, I was at a loss. I just couldn’t remember.

It is not very plausible that (32) is puzzling because we presume that the reportedly forgotten fact mattered greatly to the speaker. Why would we presume that the question of whether a murderer may kill a hypnotized aggressor in self-defence ever mattered to the speaker, when it apparently matters to hardly anyone else?232

9.2.6  A ‘Companion in Absurdity’

Even if the ‘too important’ explanation fails, the fact that (29) (‘I forgot that God exists’) strikes us as very odd may give encouragement to those who hope to provide a satisfying explanation for the puzzle of moral forgetting from a commonsense cognitivist perspective. It seems plausible that whatever it is that explains why (29) is puzzling may also explain why moral forgetting is puzzling.

This idea is plausible, I think, because moral beliefs and fundamental religious beliefs seem to be, in similar ways, different from most mundane non-normative beliefs. For instance, moral and religious beliefs are often ‘deeply held’; they amount to ‘convictions’ rather than mere beliefs with unusual frequency. A person’s fundamental moral and religious beliefs are, it may be thought, ‘part of who she is’, not (merely)

232 There is a moralized variation on the ‘too important’ explanation, according to which moral forgetting is puzzling because it is evidence of a lack of concern about moral truth, and we think that moral truths ought to matter to people. This variation faces the same counterexamples as the original. It predicts, implausibly, that we will be puzzled by cases in which someone believes a moral truth p but does not care greatly whether p, however obscure p may be.
parts of her representation of the world. It seems acceptable for a person’s moral or religious beliefs to be influenced to a degree by her personality and feelings—to a greater degree than her beliefs about geography and physics anyway. And it’s not wildly implausible to suppose that the adoption of genuine and reasonable moral and religious beliefs can involve an element of practical choice (the voluntary adoption of a stance, or the making of a commitment), rather than a simple weighing-up of evidence.

This is all very impressionistic. It is far from clear that there is any property, of the sort just mentioned, that is indeed possessed by all relevant moral and religious beliefs, and that explains our responses to (29). And the ‘companions in absurdity’ explanation (or proto-explanation) that hypothesizes that there is such a property is actually less promising than it at first appears, for any such property would have an explanatory rival. There is at least one property of beliefs in the existence of God that suffices to explain our intuitions about (29), but which is probably not shared by all of those moral beliefs whose loss through forgetting would puzzle us.

I think we normally assume that believing in God’s existence involves having faith in God. Having faith in God seems to involve, inter alia, believing that God is entitled to obedience and reverence. The distinctive features of belief in God’s existence that look as though they might explain our responses to both (29) and cases of moral forgetting are almost all distinctive features of faith in God. But of course it is possible to have a belief in God’s existence that involves no element of faith at all—for instance, one might believe in God because one thinks the universe looks designed, and yet not think we owe God anything.

The following case is designed to elicit our intuitions about the loss through forgetting of non-faith belief in the existence of God (or at least of God-like entities).

(32) If there’s a God, he doesn’t deserve our devotion. Why worship an omnipotent being who let the Holocaust happen? The problem of evil seems to rule out any God resembling that of the Abrahamic religions. Is the universe nonetheless the creation of an omnipotent intelligent being? I don’t know—I’ve forgotten.

The report of forgetting in (32) doesn’t seem to indicate a loss of faith, but it still sounds very odd. Forgetting whether or not God exists would normally involve forgetting

233 Incidentally, the hypothesis that the foregoing considerations explain why moral forgetting seems absurd is, I think, only as plausible as the ‘too important’ explanation (Section 2.5). For surely, if someone’s moral belief that \( p \) is deeply held, or is a part who she is, then it is important to her whether or not \( p \). Yet, as I have argued, the forgetting of a moral truth can seem puzzling even if that truth was apparently not important to the forgetter.
whether or not the universe was created by an omnipotent intelligent being, and (32)
suggests that forgettings of the latter sort are puzzling even when they involve no loss of
faith. Why then should we think that (29) is puzzling because, like moral beliefs,
fundamental religious beliefs are ‘deeply held’, or may reasonably be influenced by our
feelings or personalities, or involve some sort of practical commitment? We have
little reason to assume that the belief the speaker of (32) claims to have lost had any of
these properties (even when we ignore the speaker’s claim that it was lost through
forgetting).

Why does (32) sound puzzling? Perhaps because, if it were true that the universe
was created by an omnipotent intelligent being, this would be an astounding and
momentous truth, and thus a difficult one to forget. But this is only a conjecture. I don’t
think I’m obliged to provide an explanation for our responses to (32). What matters for
my purposes is, first, that it is plausible that whatever explains why (32) is puzzling also
explains why (29) is puzzling and, second, that it is not very plausible that whatever
explains why (32) is puzzling also explains why moral forgetting is puzzling. It’s hard
to think of a common distinctive feature of moral beliefs and non-faith beliefs about
omnipotent creators that might explain why the loss of both through forgetting is
puzzling.

9.2.7 Traditional Non-cognitivism

I give the label traditional non-cognitivist to a non-cognitivist who accepts what in
Chapter 7 I called the Basic Judgement-Basic Attitude thesis—who holds that basic
moral judgements do not depend on cognitive attitudes for their rationality. Perhaps the
most straightforward explanation a traditional non-cognitivist can offer for the puzzle of
moral forgetting, and the explanation I’ll focus on, is psychological. There are many
non-cognitive attitudes that plausibly cannot be lost through forgetting. The traditional
non-cognitivist can explain our puzzle by identifying moral beliefs with any of these
attitudes. This sort of explanation is not available to the cognitivist because cognitive
attitudes can be lost through forgetting (information can be forgotten).

On this account, what exactly is going on when we judge that a case of moral
forgetting seems absurd? Take a species of traditional non-cognitivism that holds that a

234 Of course, it’s not uncommon for non-faith beliefs about God’s existence to be ‘deeply held’
(e.g. in the case of passionate atheism). But a confident general assumption that they are ‘deeply
held’ seems unjustified.
moral belief is a certain sort of desire. If this is correct, when a person reports moral forgetting, she is reporting the loss through forgetting of a desire. In doing so she betrays conceptual confusion, for desires cannot be lost through forgetting. The utterance’s appearance of absurdity is explained by our grasp of the relevant concepts. To explain why we are puzzled when behaviours and feelings associated with forgetting are exhibited or experienced in connection with the loss of moral belief, a defender of this non-cognitivist theory need merely point out that such behaviours and feelings also seem incongruous in connection with the loss of desires. It would be very odd to rack one’s brains in order to re-form a lost desire, and to feel relief upon succeeding; and it is not clear that the idea of a conative analogue of memorizing and leaving reminders makes any sense.

I’ll now consider two objections to the traditional non-cognitivist explanation for the puzzle of moral forgetting. A non-cognitivist cannot avail herself of my proposed explanation if she identifies moral beliefs with non-cognitive attitudes of a sort that can be lost through forgetting. Are there any such attitudes? Among the more familiar non-cognitive attitudes, it seems intuitively clear that attitudes such as desire, preference, liking, and approval cannot be lost through forgetting. But I’m not sure about intentions. Suppose that Alice goes to the store to buy milk, but, upon arriving, is unable to remember what she went there for. She has evidently lost an intention, and it seems that this loss might be an instance of forgetting. This case is not puzzling. Moreover, I am not puzzled when I try to imagine Alice racking her brains in an effort to re-form this intention.

I’m thus not sure that non-cognitivists who, like Allan Gibbard, identify normative beliefs with intention-like attitudes can explain why moral forgetting seems puzzling. If intentions can be lost through forgetting, and if postulating that moral beliefs are intention-like provides the traditional non-cognitivist with important advantages, then the traditional non-cognitivist explanation may be in trouble. This is a real worry, which I can’t fully assuage at present. But the oft-cited advantages of identifying moral beliefs with intention-like attitudes when it comes to explaining how there can be relations of rational incoherence and disagreement among moral judgements, do not seem to be unique. A person can have incoherent combinations of

---

235 Gibbard, *Thinking How to Live.*

236 For an introductory discussion, see Schroeder, *Noncognitivism in Ethics,* 88-91.
preferences, just as she can have incoherent combinations of intentions. And, as I argued in Chapter 1, Charles Stevenson’s cases of ‘disagreement in attitude’, which are often taken to be paradigms of non-cognitive disagreement, can plausibly be interpreted as disagreements in preference rather than disagreements in intention. So I’m not convinced that the possibility that intentions can be lost through forgetting significantly undermines the traditional non-cognitivist explanation.

Let us turn to the second objection. If traditional non-cognitivism is correct, there is a species of thought and talk involving non-cognitive attitudes and their expression that, in its superficial features, quite closely resembles thought and talk involving cognitive attitudes and their expression. There is, for instance, a non-cognitive state of mind superficially resembling cognitive uncertainty. There are speech acts relating to the expression of non-cognitive attitudes that superficially resemble the information-conveying and information-requesting speech acts of asserting and questioning. There is a way of expressing non-cognitive attitudes that resembles the insincere expression of cognitive attitudes. Many more such examples could be cited.

Furthermore, to a remarkable extent, our ordinary concepts do not discriminate between these non-cognitive states of mind and communicative behaviours, and their cognitive analogues. Our concept of uncertainty applies promiscuously to both moral and non-normative uncertainty, despite the underlying psychological differences between these mental states. The same goes for our concepts of assertion, questioning, sincerity, and so on, mutatis mutandis.

Given this, and given what we know about desire-like attitudes (setting aside intentions), we might expect moral thought on a traditional non-cognitivist view to have an analogue of forgetting among its features. Forgetting is, among other things perhaps, the non-rational loss of knowledge. Desire-like attitudes can be, and seemingly are, lost over time via non-rational processes. Most traditional non-cognitivists will want to say that some of our desire-like attitudes constitute knowledge. So most traditional non-cognitivists will probably want to hold that moral knowledge can be lost non-rationally. But if ‘cognitive-seeming’ concepts such as uncertainty are as elastic, and as undiscriminating with respect to underlying psychological differences, as non-
cognitivists will want to maintain they are, why doesn’t this way of losing moral knowledge count as forgetting? Is the concept of forgetting less elastic and undiscriminating? Absent an explanation, traditional non-cognitivism, in conjunction with its typical commitments, seems to falsely predict that moral forgetting should not seem absurd.

However, I think that traditional non-cognitivists can supply such an explanation. Non-cognitivists are not committed to the view that our ‘cognitive-seeming’ concepts are undiscriminating with respect to the superficial features of our moral psychologies. And traditional non-cognitivism predicts that forgetting and its nearest moral-psychological analogue will differ superficially. For, as I have already pointed out, we are puzzled when we imagine cases in which the cessation of desire-like attitudes, apart from intentions, produces the distinctive behavioural and phenomenological symptoms of forgetting. This suggests that the non-rational loss of desire-like attitudes does not in fact characteristically generate these symptoms. These symptoms are surface features of forgetting. Traditional non-cognitivists can thus plausibly and consistently maintain that the nearest moral analogue of forgetting is just too different from paradigmatic forgetting in its surface features to be subsumed under the concept of forgetting without undermining the distinctiveness and integrity of that concept.

I have assumed in the foregoing discussion that the cognitive-seeming features of moral thought and talk fall under the same concepts as their cognitive analogues, just as they appear to. Some non-cognitivists may reject this assumption and deny the appearances. They may hold, for instance, that there is no such thing as moral uncertainty strictly speaking, or that moral uncertainty and non-normative uncertainty fall under distinct ‘uncertainty-like’ concepts. But non-cognitivists who deny the appearances must still explain them, and they must still explain why the concept of forgetting bucks the trend—why this concept does not appear to apply to its closest moral-psychological analogue. Such non-cognitivists thus face a version of the objection I have just responded to. I suspect that they can offer a similar response, but I will not attempt to show this.

Doesn’t the traditional non-cognitivist explanation face more obvious and serious problems than the two I have considered? Doesn’t the hypothesis that moral beliefs are desire-like lead to nasty semantic or meta-semantic problems? Doesn’t it make accounting for the phenomenon of moral uncertainty extremely difficult? I won’t argue otherwise here. But while these apparent defects do diminish the independent
plausibility of the traditional non-cognitivist explanation, they are inherited wholly from traditional non-cognitivism itself. And if the general problems of traditional non-cognitivism are to count against the non-cognitivist explanation, then traditional non-cognitivism’s general advantages must be allowed to count in favour of that explanation. Since traditional non-cognitivism is still, after much scrutiny, regarded as a defensible view in metaethics, its advantages (other than its success at predicting our explanandum) presumably go a fair way toward making up for its defects. My policy in this chapter, with few exceptions, has been to ignore the general advantages and disadvantages of putatively defensible metaethical theories in my evaluation of the independent plausibility of explanations for the puzzle of moral forgetting that presuppose them.\textsuperscript{239} If philosophers are right to regard these theories as defensible, this policy should not have severely distorted my evaluations, and it has allowed me to focus more narrowly on the issue of moral forgetting than I could have done otherwise.\textsuperscript{240}

The traditional non-cognitivist explanation for the puzzle of moral forgetting is the best among those I’ve considered (and the fiduciary non-cognitivist explanation that I consider in the next subsection does not improve upon it). It saves all the phenomena, without apparently generating any novel and very serious problems. Explaining our intuitions about moral forgetting is a desideratum that all metaethical theories must strive to meet, and one that deserves more attention than it has received. Traditional non-cognitivism’s success at meeting it is a hitherto under-appreciated reason to prefer traditional non-cognitivism to its cognitivist rivals.

9.2.8 \textit{Fiduciary Non-cognitivism}

It might not be clear that the non-cognitivist theory that I have defended in this thesis can, like traditional non-cognitivism, explain our explanandum. There are two apparent reasons for doubt.

First, unlike traditional non-cognitivism, fiduciary non-cognitivism rejects the Basic Judgement-Basic Attitude thesis. On the fiduciary non-cognitivist account, the

\textsuperscript{239} I arguably strayed from this policy when I compared the relative merits of non-cognitivist theories that identify moral beliefs with intention-like states and those that do not (despite the fact that theories of both sorts may be considered defensible). This was because it looked as though the former might enjoy an unusually clear and decisive advantage over the latter.

\textsuperscript{240} My objection to the version of the speaker subjectivist explanation that I considered at the end of Section 2.4 was based on the assumption that its defects were \textit{novel}, not inherited from any putatively defensible speaker subjectivist theory. This being so, I was not in a position to assume that these defects were balanced by other advantages.
non-cognitive component of a judgement of basic principle is not a basic non-cognitive attitude. It is a preference that depends for its rationality on cognitive attitudes (admissible credences) of a predominantly a priori sort. A priori beliefs can be lost through forgetting—a person can forget what single number 7×8 equals, for instance. The loss through forgetting of the cognitive attitudes on which a moralist’s voiceable preference depends would normally result in a cessation of the preference. It is not clear that such a cessation would not count as an instance of forgetting. For this reason, it is not clear that fiduciary non-cognitivism predicts that a loss of basic moral belief through forgetting would puzzle us.

I don’t think it is a problem for fiduciary non-cognitivism’s explanation of the puzzling character of moral forgetting that voiceable preferences depend on beliefs that can be lost through forgetting. For one thing, with the possible exception of intentions, non-cognitive attitudes that depend for their rationality on cognitive attitudes that can be lost through forgetting cannot apparently be lost through forgetting themselves. If a person forgets that there is beer in the fridge, and so loses her preference to go to the fridge, it is intuitively incorrect to say that this loss of preference was an instance of forgetting. She might rack her brain in an effort to recover her belief about whether there was beer in the fridge, but she would not rack her brain in an effort to recover her lost preference.

For another thing, although I have been calling the cognitive attitudes that voiceable preferences are based on ‘beliefs’, this was a rather loose way of talking. They are not genuine beliefs at all. Moralists’ admissible credences are credences concerning the consequences of following norms in the actual world, but they are based on a profoundly impoverished body of evidence, one that excludes all considerations bearing on the circumstances a moral entrusted party will find herself in at a time of action. They thus do not qualify as justified beliefs or knowledge, and consequently their non-rational cessation cannot count as forgetting. (At least this is provisionally true—I can’t be sure that the admissible credences on which simplicity desires depend cannot be lost through forgetting, because I have failed to produce an account of the provenance of simplicity desires, and producing one would require making new postulations about admissible ground-types.) So if a loss of cognitive attitudes on which a basic moral judgement depended resulted in a loss of the basic moral judgement, it could not be the case that the former loss would be an instance of forgetting, and so would ‘infect’ the latter loss in such a way as to make it an instance of forgetting too.
Here is a second apparent reason to doubt that fiduciary non-cognitivism can provide a cost-free explanation of the puzzle of moral forgetting. I have not given a full account of what moral beliefs, qua mental states, consist of for fiduciary non-cognitivists. It is possible that, on the most plausible version of fiduciary non-cognitivism, moral beliefs consist of combinations of cognitive and non-cognitive attitudes. The cognitive attitudes, I assume, would be beliefs whose truth would explain why the moral believer’s voiceable preferences made the prescription that they did about the actions in question. So if a utilitarian judged that stealing in circumstances C was wrong, this judgement would consist of a voiceable preference for conformity to norms requiring happiness-maximization, and a cognitive belief that stealing in C would not maximize happiness. (Or rather, these are the attitudes that would realize the judgement in her case; it is probably best to think of a judgement that stealing in C is wrong as the disjunctive set of pairs of voiceable preferences and cognitive beliefs that would realize the judgement.) The cognitive component of a basic moral belief would, I assume, be an analytic triviality. For instance, a utilitarian’s belief that happiness-maximization is necessarily required would consist of a voiceable preference for conformity to norms requiring happiness-maximization, and a cognitive belief that maximizing happiness is necessarily happiness-maximizing.

Suppose that, for the fiduciary non-cognitivist, moral judgements are composed of cognitive and non-cognitive attitudes in the way I have described. Cognitive attitudes can be lost through forgetting, and if a cessation of moral belief resulted from the loss through forgetting of its cognitive component, it is not clear that this cessation would not be an instance of forgetting.

My response to this problem is predictable, because it is not unobvious, and because it echoes a point I made on behalf of the intuition explanation in 9.2.3.241 It is not desirable that fiduciary non-cognitivism should predict that all prima facie cases of moral forgetting seem absurd, in part because these include case of LRNF, and LRNF does not seem absurd in general. If fiduciary non-cognitivism predicts that the loss through forgetting of the contingent-belief component of a derived moral belief can result in the loss through forgetting of the moral belief, this is not obviously any cost to fiduciary non-cognitivism. It would be a problem if fiduciary non-cognitivism predicted that we would not be puzzled by cases in which a moralist lost through forgetting the

241 The point concerned whether or not ‘C’ sufficed to predict our explanandum.
cognitive component of a *basic* moral belief, and reported having lost the basic moral belief through forgetting. This would contradict our explanandum. But the version of fiduciary non-cognitivism under consideration does not make this prediction. It says that the cognitive component of a basic moral belief is a very obvious analytic belief. Thus the case ought to puzzle us merely in virtue of the fact that it asks us to imagine that someone might lose such a belief through forgetting.
Conclusion

If we are prepared to make certain assumptions and overlook a few omissions, fiduciary non-cognitivism broadly meets the nine desiderata that I listed in the Introduction (namely, ‘Disagreement’; ‘Justification, criticism, and demands for grounds’; ‘Moral uncertainty’; ‘Moral reasoning and changes of motivation’; ‘The existence of morality’; ‘Ubiquity of moral thought and talk’; ‘Rule-like character of moral requirements’; ‘Moral forgetting’; ‘Moral approval and disapproval’).

The theory explains the robustness of moral disagreement—though I have only shown this for disagreements of basic principle, and not in cases of purely hypothetical or past-tense disagreement—by positing that moral judgements involve voiceable preferences of parties to a dual-order fully connected network, and by giving an account of moral judgement which ensures that the voiceable preferences involved in permission and prohibition judgements about the same action are mutually unsatisfiable.

It explains the use of moral judgements to justify and criticize, and explains our entitlement to demand that people give the grounds for their moral judgements. Parties to dual-order fully connected networks can, in general, voice preferences in order to defend or object to actions. They have discursive entitlements to demand that other members should explain why they have the voiceable preferences they do. More specifically, members of the moral trust network have similar entitlements to demand that other moralists reveal the ‘norm-principles’ that explain their judgements. The explanations will be less satisfying to other moralists if the norm-principles lack generality, or would not plausibly be compatible with any simple comprehensive norm-partition that also entailed the moralist’s previously avowed norms.

Fiduciary non-cognitivism can account for uncertainty about basic moral principles. Moral indecision is a state of having two or more best-preferred comprehensive norm-partition between which one’s expected weighted utilities qua moralist fail to discriminate. A moralist has less than perfect confidence in a norm if the expected weighted utility for her (behind the ‘veil of ignorance’) of conforming to her best-preferred comprehensive norm-partition among those entailing the norm is lower than a certain value N. N is the value such that, if she were perfectly rational and logically omniscient, conformity to her best-preferred comprehensive norm-partition
would have an expected weighted utility of $N$ for her. It is often rational to act on a moral judgement which one holds with imperfect confidence, according to fiduciary non-cognitivism, because doing so will normally be the best way to satisfy one’s discretional desires qua moralist.

Unlike most non-cognitivist theories, fiduciary non-cognitivism explains the changes in moral motivation that not uncommonly result from substantive moral reasoning without committing to the thesis (‘Heft’) that higher-order non-cognitive attitudes reliably cause changes in lower-order ones. Moral reasoning is instrumental reasoning affecting voiceable preferences. When it produces new voiceable preferences, our discretional desires reliably motivate us to act on them.

Fiduciary non-cognitivism makes the existence of morality less surprising than other non-cognitivist theories do. It posits that our moral practices and thoughts are similar in kind to other more localized practices and thoughts that are plausibly found in clanish families and communities for instance; and holds that the same sorts of factors that explain the existence of these localized practices explain the existence of the moral trust network. The theory does have significant commitments in evolutionary biology, but these appear to be somewhat smaller than those of rival theories like Gibbard’s, because it explains many facts about moral thought and talk by pointing out the instrumental advantages of the ‘institution’ of morality, rather than by giving them a biological explanation. My failure to explain moralists’ simplicity desires is admittedly a serious problem for my theory, but it is not clear to me that rival non-cognitivist theories can provide a good (presumably evolutionary) explanation of the virtuousness of simplicity in moral theorizing either.

I have gamely striven to explain the ubiquity of moral thought and talk by arguing that, in certain environments, there is a way in which humans might have evolved dispositions to extend their trust and commitments to other humans almost indiscriminately, in the way that the existence of an all-encompassing moral trust network requires. More convincingly, I have also argued that we must posit something like the moral trust network if we are to give a wholly satisfactory explanation of the phenomena of moral approval and disapproval.

According to fiduciary non-cognitivism, moral requirements are apparently rule-like in character because abuses of trust by parties to the moral network will be easier to detect if moralists only have voices in moralists’ choices of which rules or norms to follow, and are somehow induced to prefer that moralists should follow simpler rather
than more complex norms. The simplifying influence on moralists’ preferred norms is a
weighted desire, possessed by almost all moralists, that moralists should follow simpler
norms. This is another feature of my theory that is undermined by my failure to explain
why moralists would have such desires.

Like most traditional non-cognitivist theories, fiduciary non-cognitivism correctly
predicts that we will be puzzled by cases that invite us to think that someone has lost a
basic moral judgement through forgetting. Such judgements, according to fiduciary non-
cognitivism consist in or are based on attitudes whose loss through forgetting is either
impossible (desire-like attitudes) or would be very surprising (admissible credences,
trivial analytic beliefs).

As I more or less noted above, fiduciary non-cognitivism explains the phenomena
of moral approval and disapproval in light of Philippa Foot’s observation that
approbation seems to require a certain kind of social context, by postulating the
existence of a trust network linking all moralists.

These virtues notwithstanding, my theory and my argument for it are deficient in
crucial respects. Chief among these are my aforementioned failure to explain why
moralists have simplicity desires, and my commitment to the possibility and
significance for rational decision-making of infinitesimal (or perhaps just near-
infinitesimal) credences. Though I have argued that fiduciary non-cognitivism’s
commitments in evolutionary biology are smaller than is typical for a non-cognitivist
theory, they are still weighty. Some of my evolutionary conjectures seem far-fetched
from the philosopher’s armchair, and I haven’t yet attempted to assess them empirically.

I have not established whether the trust relationships on which fiduciary non-
cognitivism relies can exist in the absence of normative judgements of the sort that the
theory seeks to explain; nor have I established whether, if they cannot, relying instead
on some surrogate for trust (say, ‘reliance’) in my theorizing will result in an adequate
theory. I have not examined the possibility and implications of public defection from the
moral trust network; such an inquiry might reveal serious difficulties for my theory.
Fiduciary non-cognitivism about morality will be much less plausible if a fiduciary non-
cognitivist construal of other kinds of normative and evaluative thought and talk turns
out to be implausible, and I have not attempted to show that this is not the case.

My endorsement of the Disparity Hypothesis in 8.2.1 raises a set of problems for
my theory that I haven’t properly acknowledged yet, nor indeed properly fathomed. The
Disparity Hypothesis, it will be recalled, holds that as a general matter, an entrusted
party is obliged to perform the action with the highest expected weighted utility for his partner relative to his (the entrusted party’s) evidence, not to accommodate her voiceable preferences; and that he is entitled to resist doing so only if it is to perform the action with the highest expected weighted utility for him relative to his evidence, not the action that satisfies his voiceable preference. This makes pressing a number of questions concerning the relationship between an entrusted party’s voiceable preferences and his estimations of what has the highest expected weighted utility for him relative to his evidence. When an entrusted party takes advantage of his discretion to resist his partner’s demands, should we expect him to choose the option that he voiceably prefers, or the option such that he a believes that it has the highest expected weighted utility for him relative to his evidence? Can these two options differ if he is rational? Would the second possibility have the result that the practical deliberation of moralists involves an implausible amount of introspection and theorizing about their own psychologies?

Moral reasoning is a difficult enterprise, and a moralist will often have a very low confidence that the option that had the highest expected weighted utility for him by his lights was the option with the highest expected weighted utility for him relative to his evidence. Might such a moralist be too afraid to act on his norms for fear that he was exceeding his discretion? I have ideas as to how to address these questions—some involve the liberal postulation of what in 8.2.1 I called ‘excuses’ for acting on false estimations of what would best satisfy weighted desires—but of course I have not defended them here.

These are all problems for my theory specifically. Among the phenomena that metaethicists in general seek to explain, there are a host that I have not addressed in this work, and it remains to be seen whether fiduciary non-cognitivism can account for them. I have not shown the fiduciary non-cognitivism can supply an adequate account of moral language or deductive reasoning. I have not provided a full account of what simple, prescriptive moral judgements are, and have said nothing about other kinds of moral judgement (e.g. evaluative judgement). I have not attempted to explain the phenomena of forgiveness, praise, and blame.

These deficiencies and omissions all suggest directions for future research. And of course they are important to an evaluation of my theory. I do not think they warrant its rejection. As I suggested in my introduction, the profusion of evident problems facing my theory is partly explained by its degree of detail and explicitness, and by my efforts to provide a deep and unified explanation for almost all of its postulations. I do not
believe that any rival non-cognitivist moral psychology is similarly detailed and explicit, or has been so thoroughly explained. It is quite possible that attempts to explain and make precise these rival theories to the same degree will reveal a similar profusion of problems. Until this work is done, the greater overall success of fiduciary non-cognitivism at meeting my nine desiderata argues for the superiority of that theory to its non-cognitivist rivals.
Appendices

Appendix A. Debt

I said in my characterization of entrusted concern relationships in 1.2.2 that an entrusted party strains his partner’s trust by refusing to yield and relieves strain by yielding, and that if he places too much strain on her trust over time, he will eventually violate it. This means that the more strain an entrusted party has accumulated, the more solicitous he must be in accommodating his partner’s preferences. Given this, and given my claim that yielding by an entrusting party allows her partner to defy her accommodation-obliging preference without straining her trust, why would an entrusting party ever want to yield? Surely it would serve her interests better to let her partner strain her trust without giving him her blessing. That way, she could expect greater deference from him in the future.

However, I suggest there are ways in which the entrusting party can yield that involve no such sacrifice of influence. What I am calling ‘strain’ need not be the only constraint on the extent to which an entrusted party can resist his partner’s preferences without violating her trust. An entrusted party can, I posit, possess a thing that I will call debt, which is like strain in that, the more of it the entrusted party accumulates, the less scope he has to resist his partner’s demands without violating her trust; but unlike strain in that the entrusted party’s level of debt is the outcome of negotiation and agreement by the parties. Once the parties have reached a common understanding of what the entrusted party owes the entrusting party, then any strain the entrusted party has put on his partner’s trust ceases. Their trust relationship is then in better shape. Respectively, entrusted and entrusting party have a clearer idea of what is expected, and what to expect, of the entrusted party.

An entrusted party who intends to resist his partner’s preference might communicate to her an offer to acknowledge that he would thereby acquire a certain amount of debt to her, if she will assent to his resisting her preference. His primary motivation for doing this would normally be that he wants to preserve his relationship with his partner, and so wants to avoid uncertainty about what is expected of him. Under these circumstances, yielding by the entrusting party is a way for her to at once assent to his resisting her preferences, and to acknowledge that this is indeed the amount
of debt that he will accrue by doing so. In fact, acknowledging the level of debt that the entrusted party would accrue by resisting her preferences would, given that he has already conditionally recognized this debt himself, be equivalent to yielding. For she would bring it about that his resisting her preference would not strain her trust, thereby eliminating her trust-based entitlement that he should accommodate her preference in this instance. Once this mutual conditional acknowledgement had been reached, the entrusted party could choose to either accommodate her preference or resist it and incur a debt, and the entrusting party would not have a voice in this decision. Her situation is indistinguishable from the situation an entrusting party is in if she has yielded.

Appendix B. ‘Excuses’ to Resist Preferences

I argued in 3.1.4 that the common entrusted party in a uniform weighted desires network does not significantly strain his partners’ trust by acting on his discretion-licencing preferences in defiance of their preferences. This is how things would need to be to guarantee that situations in which he has no choice but to violate trust never eventuate, on the following assumption: that an entrusted party B strains an entrusting party A’s trust equally when he exercises his discretion to resist A’s preference, regardless of his circumstances.

I am now going to propose that we drop this assumption. The terms of an entrusted concern relationship may place distinctive conditions on the circumstances under which resisting an accommodation-obliging preference strains A’s trust. I propose that B may have excuses to resist such a preference, and that, if so, he does not strain A’s trust by so resisting, or he strains it to a lesser degree. More specifically, I propose that, often, the common entrusted party in a uniform weighted desires network (and in many similar networks featuring multiple entrusting parties and a common entrusted party, and in which the Uniform Weighted Desire-Types solution is in effect) has an excuse to resist an entrusting party A1’s preference if there is some entrusting party A2 in the network such that A2’s preference opposes A1’s and A2 has not waived. Call this a conflict excuse—the idea behind this label being that an entrusted party may resist an entrusting party’s preference without straining her trust if this preference ‘conflicts’ with that of another entrusting party.

Having this conflict excuse should affect not the entrusted party’s behaviour in any way that is novel to us. The excuse merely permits him to act on his discretion-licencing preference in defiance of A1’s preference without straining A1’s trust—it
does not, for instance, permit him to yield to A1’s preference without violating the trust of any other entrusting parties whose preferences oppose A1’s. According to my characterization of uniform weighted desires networks, acting on his discretion-licencing preferences is pretty much all that the common entrusted party in a uniform weighted desires network does anyway, as such (at least if it is large). So why postulate that common entrusted parties can have conflict excuses? The reason is that this allows us to posit that, in relevant networks, absent a conflict excuse, a common entrusted party will significantly strain his partners’ trust when he acts on his own preference in defiance of theirs. This means that when all of his partners who have not yielded or waived are unanimous in their preferences, a common entrusted party cannot resist these preferences without straining those partners’ trust. This arrangement will be advantageous in many uniform weighted desires networks and other similar networks, especially if they are small, because it will give the entrusting parties some power to thwart the entrusted party when his preferred options would frustrate the weighted desires of everyone else in the network (bracketing waiving parties). So assuming that conflict excuses are a possible feature of entrusted concern relationships, they ought to be common in such networks.

Postulating the possibility of excuses is not, I think, very costly. I suggested in Appendix A that ‘strain’ exists only absent an understanding between an entrusted party and his partner about what the former owes the latter. It can be replaced with ‘debt’ of an agreed magnitude, through the voluntary concessions, acknowledgements, and indulgences of the parties. I am now merely postulating that the terms of entrusted concern relationships might include rules governing what sorts of understandings may be reached, concerning the entrusted party’s debt, consequent upon his resisting his partner’s preferences in certain circumstances. We can think of a rule granting an entrusted party conflict excuses as a rule prescribing the immediate substitution of a debt increase of zero for any strain he causes when his partners’ preferences are not unanimous.

My postulation of the possibility of conflict excuses is theoretically motivated. As I will explain in Appendix C, it allows me to more credibly postulate that networks similar to uniform weighted desires networks would exist in quite small groups. This in turn allows me to give a better explanation of our intuitions about third-party disagreement in cases like Football (Parents I).
Appendix C.  

Explaining Third-Party Disagreement in Football (Parents I)

I’ll consider here why we might tacitly attribute to John and his parents in Football (Parents I) a network of trust relationships that would permit third-party disagreement about what John should do. I will assume that John’s is a stereotypical nuclear family whose members all live under one roof, and that John’s parents are married. Something like the base conditions for a fully connected network obtain here. John’s family members would have common desires for the promotion of the family’s welfare, status, and shared values (or, if the word ‘values’ is inappropriate here, their ‘more distinctive shared concerns’). These common desires (call them W) might vary in strength among the members of John’s family. And John is not the only member of his family who could satisfy or frustrate these common desires through his actions. This means that John and his parents (and his siblings, if he has any) would have some incentive to form a fully connected network. Let’s suppose for the moment that they would form such a network, and consider what features it would have, and whether we should expect the members to be capable of having third-party disagreements.

In addition to their first-order fully connected network, it would be advantageous for John’s family if the preferences they voiced were subject to more rather than less scrutiny, which a second-order fully connected network, entitling each party to a voice in the others’ waiving choices, would facilitate. So they would have some incentive to form a second-order network. This would be somewhat desirable from the point of view of each member. If John’s father urged John to stay on the football team by voicing a preference that he should do so, John’s mother would probably like to be able to voice a contrary preference, and to demand to know the reasoning behind her husband’s preference, without being ignored by him. Enjoying this entitlement would more than compensate John’s mother for the chore of having to explain her voiced preferences on demand.

At least under some circumstances, we would intuitively expect the entrusted parties in this network to significantly strain their partners’ trust by resisting their preferences. So these entrusted parties would, I propose, have conflict excuses for resisting trust (see Appendix B), to ensure that they would never find themselves in situations in which they could not avoid violating trust as a result of accumulated strain.

The relationships shared by the members of these two fully connected networks as such need not be the only entrusted concern relationships they share. I’m inclined to think that there would be different and more complex entrusted concern relationships, of
a more personal nature, between pairs of family members, whose terms would not be wholly determined by the requirements of any larger trust network. John’s parents are presumably romantically attached and sexually intimate to some degree, and it seems possible that they would be able to apply trust-based pressures to each other in these roles, by voicing preferences based on desire-types other than those that were weighted for them qua John’s partner. This is compatible with the possibility that a fully connected network would exist among the members of John’s family, in which everyone’s weighted desire-types were identical.

Despite the advantages I have mentioned for John’s family of forming a dual-order fully connected network, the hypothesis that John’s family members would do so lacks plausibility. If it were true, John would presumably have an influence in his parents’ decisions equal to John’s parents’ influence in his (as long as we are considering only the sort of influence wielded by entrusting parties, and not the capacity to make authoritative commands, which presumably only John’s parents have). This seems prima facie implausible given the presumed disparity of power between John and his parents, and at any rate seems contrary to our ordinary assumptions about children’s relationships with their parents.

John’s preference might decide whether or not his father could yield to his wife’s preference in some matter even if he (the father) had a strong pre-trust desire to yield; it might decide whether John’s father would strain his wife’s trust at all by resisting her preference. Imagine John’s father responding to a demand of his wife by saying, ‘I would do it, but the boy doesn’t want me to.’ This would seem unusual. It would seem unusual too if John’s father thought he had to get John on his side before urging John’s mother to make some choice. But this would make perfect sense if John’s family had formed a fully connected network, and unanimity of preferences were required among the partners of a common entrusted party before the latter could yield or face significant pressure to yield.

However, the Four Relationships Hypothesis does not require that the common entrusted party whose decisions are in dispute be an entrusting party himself. So we can drop the assumption that John has a voice in his parents’ decisions—or one that is on a par with their voices in his decisions anyway—and still potentially explain the intuition.

However, the decisions in which these more personal trust relationships granted entrusting parties voices might have to be circumscribed in certain ways to prevent interference with the dual-order fully connected network.
of disagreement elicited by *Football (Parents I)*. Suppose that we make this move. I contend that it is still plausible that the trust relationships necessary for third-party disagreement would exist amongst John and his family. It is still plausible that a network of trust relationships would exist that functioned to encourage the satisfaction and inhibit the frustration of the family’s common desires. It would merely be a less ‘fully connected’ one, for it would omit to give John a voice in his parents’ decisions.

I hypothesize that there would be a special set of entrusted concern relationships N among John’s family members whose function was to promote the satisfaction of their shared desires W. I’ll label the relationships in N using the following notation: E(A, B, nth, N) refers to the nth-order entrusted concern relationship in N, in which A is the entrusting and B the entrusted party. Qua members of N, John and his parents would share the following trust relationships: E(Mother, Father, 1st, N), E(Father, Mother, 1st, N), E(Mother, John, 1st, N), E(Father, John, 1st, N), E(Mother, Father, 2nd, N), and E(Father, Mother, 2nd, N). Notice that John is not the entrusting party in any of these relationships.

Before I specify why I think these relationships would exist, let me comment on two distinctive features that I think N would have. The weighted desires of all the parties in these relationships are, I hypothesize, all W-type desires, as we would expect given the function of N—but with one exception. The set of John’s weighted desire-types in E(Mother, John, 1st, N), E(Father, John, 1st, N) would probably be larger than W. (The set would, of course, be the same in both of these relationships—this is needed to rule out (2)-type situations.) More precisely, it would probably be a proper superset of W, including desire-types that were more self-regarding than those in W. I propose this to account for the intuitive fact that John might cite his enjoyment of football as a ground for his discretion-licencing preference to play football, as he does in *Football (Mother and Son)*. It is not clear that a desire for John’s enjoyment would be among the family’s shared ends.

This would make sense given John’s situation. Because John’s reasons for being part of N are different from those of his parents, it is not important that his weighted desire-types should be identical to those of the other members. Unlike his parents, John is not a member of N chiefly in order to secure the satisfaction of desires he holds in common with the rest of his family. Rather he is a member because his parents have power over him and can make life difficult for him if he is not willingly responsive to their desires. He does not commit to satisfying any particular set of desires (his or
theirs) to ensure that they reciprocate, and so does not need to commit to satisfying desires of the same types that they commit to satisfy.

There is admittedly a minor drawback in John’s having a peculiar set of weighted desire-types. Uniformity of weighted-desire types facilitates, and provides occasions for, the scrutiny of preferences. For instance, it is somewhat desirable that John’s father should have the same weighted desire-types as his wife does in E(Mother, Father, 1st, N), because this means that, when his wife gives the reasoning behind her accommodation-obliging preference, there is a stronger presumption that the same reasoning will give him a preference for the same option. This means that when she voices a preference, she challenges him to find a flaw in her reasoning, with the result that errors in the reasoning of both parties will potentially be exposed. The same is true, mutatis mutandis, of all entrusted concern relationships among the members of N. Hence it would be desirable for John’s mother to have the same weighted desire-types as John does in E(Mother, John, 1st, N), for instance. John’s having a peculiar set of weighted desire-types means that the reasoning behind his mother’s accommodation-obliging preferences is relatively unlikely to apply to his discretion-licensing preferences, and expressed disagreements between John and his mother are relatively unlikely to expose bad reasoning. But this is only a small disadvantage.

Let’s now turn to the second distinctive feature of N. In general, because N is presumably very small, and because an entrusted party who resists accommodating the unanimous preferences of the other entrusting parties in N significantly strains their trust, members of N who want to change other members’ behaviour will not have to resort to trying to change the reasoning behind their discretion-licensing preferences. They will often be able to exploit one another’s accommodative desires (i.e. their desires to avoid straining trust) as a means of influencing behaviour. Achieving unanimity of preferences among the partners to a common entrusted party who have not waived will often be a feasible goal here. It will be feasible because the set of entrusting parties in N is so small. Indeed, John’s parents are probably the only entrusting parties in N. They are probably also the only parties in N who share second-order entrusted concern relationships. This means that their second-order relationships will be of a ‘personal’ sort, in which resisting preferences results in strain, and in which entrusted

---

243 If there are other children in the family, they will probably only participate in N as entrusted parties, as John does.
parties are free to yield without having to worry about whether their partners are ‘unanimous’. If this is right, and if John’s parents disagree about what John should do, the parent X who disagreed with John would often be able to induce the parent Y who ‘agreed’ with John to waive, by exploiting Y’s desire not to strain X’s trust, and perhaps by signalling her willingness to be similarly acquiescent herself in future waiving choices if Y would waive in the present case. They could thus use their voices in their second-order relationships to end their disagreement concerning what John should do, so that one of them could make effective demands of him. (Indeed we should pretheoretically expect them to do something like this: disagreeing parents typically prefer to reach a common position concerning their child’s decisions, so as to avoid making incompatible demands of the child.) Overall, we should expect the entrusted concern relationships among the members of N to allow for a lot more practical negotiation than we would find in large fully connected networks, as opposed to mere scrutiny of reasoning.

Let me now summarize why I think that the relationships that I have posited would in fact exist. John’s parents would share two first-order entrusted concern relationships \(E\langle \text{Mother, Father, 1st, N} \rangle\) and \(E\langle \text{Father, Mother, 1st, N} \rangle\) in which their \(W\)-type desires were weighted because they have a common interest in satisfying their \(W\)-type desires, and each would be willing to commit to give greater weight to \(W\)-type desires in her decisions if the other would reciprocate. They would entrust to John their concerns that he should satisfy their \(W\)-type desires, and would use their power over him to induce him to commit to do so (hence \(E\langle \text{Mother, John, 1st, N} \rangle\) and \(E\langle \text{Father, John, 1st, N} \rangle\)). The sets of John’s weighted desire-types in these relationships might not equal \(W\), but they would be the same in both relationships. John’s parents would also share two second-order entrusted concern relationships entitling each to a voice in the other’s waiving choices qua John’s partner. They would do this partly so that they would have discursive entitlements that would aid each in scrutinizing the demands the other made of John, but also because it would frequently be especially advantageous to one party if she could get the other party to waive, and these relationships would facilitate reciprocal waiving: each parent X might be willing to waive in a decision of John’s that mattered relatively little to X, in the expectation that the other parent would waive in a future decision that mattered more to X (hence \(E\langle \text{Mother, Father, 2nd, N} \rangle\) and \(E\langle \text{Father, Mother, 2nd, N} \rangle\)). I hypothesize that we ‘tacitly’ assume that these
relationships exist when we are presented with Football (Parents I), and that this explains the appearance that John’s parents disagree.

Appendix D. Explaining Third-Party Disagreement in Penny’s Case

This appendix assumes knowledge of Chapters 1-3 and Appendix B, and appeals to the idea of non-cognitive disagreement about conditional decisions, the possibility of which I consider in 4.2.

As I observed in 3.2.2, explaining why the conditions for third-party disagreement would be met in Marriage (Imperatives) seems a less straightforward matter than explaining why they would be met in Photographs. Simply positing a dual-order fully connected network among Penny’s family members seems unsatisfactory. Something needs to be said about how such a network—or whatever network we posit—would reflect the presumed disproportionate power that Penny’s mother has over Penny, or if it did not accommodate this power, why she and Penny would nonetheless be members of it.

One way to avoid predicting that everyone in Penny’s extended family would have an equally influential voice in Penny’s choice of spouse would be to make the following five postulations. First, Penny can significantly strain her mother’s trust by resisting her preferences, but has conflict excuses for resisting them. Second, the fully connected network does not give members a voice in Penny’s choice of what to do when this involves going against her mother’s wishes. Third, the network nonetheless gives members voices in Penny’s mother’s waiving choices qua entrusting party to Penny. Fourth, any preferences voiced by the members of the network, apart from Penny’s mother, concerning what Penny should do are, insofar as they are directed at Penny, tacitly conditional. They concern what Penny should do given that this doesn’t involve going against her mother.

Before continuing, I should clarify that these are the only voiced preferences that we need to posit would be tacitly conditional in this way. Insofar as they are directed at Penny’s mother, these members’ voiced preferences concern her waiving choices qua first-order entrusting party to Penny—not in the matter of what Penny should do when this does not involve going against her mother, but simply in the matter of what Penny should do. In the first-order entrusted concern relationship between Penny and her mother in which Penny is the entrusted party, both parties have voices in the matter of
what Penny should do, and not in her choices of what to do when this does not involve going against her mother’s wishes.

The final postulate is that the admissible ground-types for all parties in the network would rule as inadmissible any proposition a positive credence in which could make it rational for her to hold, for all φ, a voiceable preference concerning whether Penny should φ that differed from her preference concerning whether Penny should φ given that Penny would defy her mother’s wishes by φ-ing. I suggest that they would do this by ruling facts about whether Penny’s mother has yielded fundamentally inadmissible. This would not be an unusual sort of arrangement: facts concerning the actions that parties to a fully connected network might perform as such often need to be inadmissible in such networks if instances of the Reaction Problem are to be avoided.

If we suppose that Penny’s aunt sincerely voices a preference for Penny to marry Mr. Lloyd, and Penny’s mother sincerely voices a preference for Penny to marry Mr. Toff, these five postulations have the following implications. Penny’s aunt and mother disagree about whether one of them should waive (depending on Penny’s preference qua first-order entrusted party in each relationship); Penny either disagrees with her mother about whom she should marry, or disagrees with her aunt about whom she should marry, *if this doesn’t involve going against her mother’s preference*; she cannot disagree with both of them. At least this is true if we assume that everyone has relevant voiceable preferences.

It is a further question whether Penny’s aunt and mother are in third-party disagreement—that is, whether they disagree about something that Penny should do. I suggest that they at least disagree on the conditional question of whom Penny should marry if it doesn’t involve going against her mother’s wishes. Penny’s mother prefers that Penny should marry Toff unconditionally—i.e. whether or not Penny would go against her mother’s wishes by doing so—so she prefers that Penny should marry Toff when this means going against her mother’s wishes. This preference and Penny’s aunt’s are mutually unsatisfiable, and the requisite trust relationships exist between them and Penny for this to put them in third-party disagreement according to the Four Relationships Hypothesis.

Having to posit that the aunt and mother only disagree on a conditional question—the question of whether Penny should marry Lloyd if doing so does not involve defying her mother—seems to be a cost. For one thing, it is costly in general to posit that someone is implicitly saying something conditional when her utterance is, on
the surface, unconditional. For another, it is costly to posit the an arrangement of trust relationships as elaborate as that I have described among Penny and her relatives. Finally, my postulation does not obviously lend itself to a plausible construal of *Marriage (Imperatives)*, where the aunt clearly knows that Penny would defy her mother by marrying Lloyd but tries to get her to do it anyway.

However, that case, which has the aunt and mother barking conflicting commands at Penny, is quite silly and stylized. In a more natural version of the case the following would not seem implausible: that the aunt is not really urging Penny to defy her mother, but rather simultaneously urging Penny to marry Lloyd if her mother does not oppose it, and urging Penny’s mother not to oppose it. This proposition is consistent with my account of the trust relationships among Penny and her relatives. Moreover, I think that this account provides at least a defensible explanation of our intuitions about *Marriage (Imperatives)*, fanciful as that case is. For it is somewhat plausible that Penny’s aunt is getting carried away and exceeding propriety by demanding that Penny defy her mother. The interpretation of the case I am defending allows for the possibility that the aunt’s demand is being presumptuous while still correctly predicting disagreement between mother and aunt.

How does the arrangement of trust networks in Penny’s family that I have hypothesized better reflect the underlying power relationships among the members? It at least means that the mother’s influence on her daughter’s choice will never have to compete with the influence of other first-order entrusting parties in the network. As soon as she has a (rational) preference in Penny’s choice of whom to marry, other extended family members lose their voices in the decision. This means that Penny’s mother’s demands of her daughter are more likely to be efficacious, especially since Penny can significantly strain her mother’s trust by failing to accommodate her preferences. It doesn’t mean that Penny’s mother can make special, unusually stringent demands of her daughter: Penny’s mother’s preferences must be based on the same weighted desires that all her preferences are based on in the first-order network.²⁴⁴

Of course, I am not wedded to this interpretation; it is just one of a disjunction of possibilities that I am committed to. If we are prepared to accept that Penny’s aunt and

²⁴⁴ There would plausibly be a more personal entrusted concern relationship between Penny and her mother that was not part of any fully connected network, and that would allow Penny’s mother to voice preferences based on different desires. But this relationship and the fully connected network among Penny’s extended family would presumably give Penny’s mother voices in different sets of decisions, among those facing her daughter.
mother have roughly equal voices in Penny’s decision, we can get by with positing that the two disagree in virtue of their membership of a simple dual-order fully connected network. We can then explain the disagreement between aunt and mother without positing that it is about a conditional decision of Penny’s. We can also, in a weak and limited way, allow that Penny’s mother has a more influential voice in Penny’s decision than her aunt does: we can posit that Penny can significantly strain her mother’s trust (and not her aunt’s) by not yielding to her preferences, absent conflict excuses.

Appendix E. Third-Party Disagreement and the Disparity Hypothesis

In this appendix I shall try to show how fiduciary non-cognitivism’s explanation of moral disagreement can be made compatible with the Disparity Hypothesis (8.2.1). As far as I can see, I will have accomplished this once I have made certain revisions to my account of waiving, and shown that my theory of third-party disagreement plausibly has the virtues I have claimed for it once I have made these revisions.

The Disparity Hypothesis requires a revision of my account of waiving because waiving affects entrusted parties’ obligations, and the Disparity Hypothesis concerns the nature of entrusted parties’ obligations. I can’t ignore these required revisions (as I shall ignore the revisions the Disparity Hypothesis requires for my account of yielding) if I want to explain moral disagreement, because my explanation of third-party disagreement in general relies heavily on the idea of waiving.

The necessary revision is fairly straightforward. As I posited in 3.3.3, for all φ, an entrusting party A can waive her demand that her partner B should φ just in case A voiceably prefers that B should φ. On my revised account, as long as she continues in her waiving stance,

(i) if B lacks a voiceable preference to not-φ, and B not-φs, and φ-ing is the action with the highest expected weighted utility for A relative to B’s evidence, then B does not strain A’s trust by not-φ-ing (even if B does not know that A has waived); and

(ii) if a suitable occasion arises, and if A thinks that it will make a difference to whether or not B φs, A must communicate to B that she has waived with respect to B’s φ-ing.

Given this account of waiving, we should expect to have an intuition of third-party disagreement in cases in which Conditions One to Five hold. If S, T, and U share the relationships and have the voiceable preferences they must have, according to the Four Relationships Hypothesis, for S and T to disagree about whether U should φ, then a
state of affairs will exist that deserves to be called a non-cognitive disagreement between S and T about whether U should φ.

The Four Relationships Hypothesis crucially requires a guarantee of ‘harmony’—a correlation between the voiceable preferences of each party to a third-party disagreement in each other’s waiving choices, and their voiceable preferences concerning the entrusted party’s action. The changes I have made to my account of waiving do not seem to undermine the possibility of guaranteeing harmony, or so I shall now argue.

Before we take into account constraints on their admissible ground-types, if S prefers that U should φ, and T prefers that U should not-φ, then each will tend to prefer that the other should waive, for this will tend to make it more likely that U will do as she prefers. If T waives, for instance, then, whenever T thinks that it will make a difference to whether U φs and has a suitable opportunity, T must let U know that he has waived. T’s telling U that she has waived will generally make a difference to whether U φs only when U does not disprefer T’s preferred option (not-φ), and in that case it will tend to increase the likelihood that U φs. This is because, if U knows that he would not violate T’s trust by φ-ing if φ-ing did not have the highest expected weighted utility for U relative to U’s evidence, his discretionary desires qua T’s partner will not urge him to avoid φ. So S will tend to prefer that T should waive. T will tend to prefer not to waive for the same reason that S prefers that he should waive—it will raise the probability that U will frustrate his voiceable preference.

Constraints on admissible ground-types will turn these tendencies into guarantees of harmony. Any proposition p such that S’s belief in p might explain why S would, despite preferring that U should φ, rationally prefer that T not waive his preference that U should not-φ, would be made inadmissible for S by the terms of her trust relationships. A similar pruning of T’s admissible ground-types will guarantee that T prefers not to waive.

Part of my explanation in 2.2.2 for the appearance that a third-party non-cognitive disagreement is a disagreement about the actions of the third party was roughly as follows. According to the Four Relationships Hypothesis, if S and T disagree about whether U should φ, there is necessarily at least one disagreement between them concerning whether one of them should waive qua entrusting party to U. Whose waving choice(s) this concerned depended on what U preferred to do. I argued that since it was their incompatible preferences about U’s choice that guaranteed that they shared at least
one such disagreement (and since their disagreeing thoughts were ones to which U was obliged to give consideration), their situation deserved to be called a disagreement about what S should do.

I have not yet acknowledged it, but my postulation in 3.3.3 that an entrusting party may enter a waiving stance concerning an entrusted party’s decision even if the entrusted party’s preference opposes hers, requires me to give a slightly different explanation of why S and T’s disagreement is about the third party’s choice. An explanation that is compatible with the 3.3.3 postulation cannot appeal to the idea the parties are robustly in non-third-party disagreement only because of their opposing preferences concerning the third party’s actions. My new account of waiving inherits this feature of the 3.3.3 version, so I still need to provide an alternative explanation.

The 3.3.3 postulation means that if S and T have opposing preferences concerning what U should do, and the network of trust relationships required by the Four Relationships Hypothesis obtains among the trio, then there will always be two disagreements between S and T—a disagreement about whether S is to waive, and a disagreement about whether T is to waive. To be sure, if (e.g.) S and U prefer the same option, then the disagreement about whether S should waive will be more pressing and practically significant than the disagreement about whether T should waive, because it is the disagreement whose outcome has a more immediate prospect of changing U’s behaviour. But both disagreements will exist. S will still want T to waive, because this means that if U comes to prefer T’s preferred option, U may still choose the option with the highest expected weighted utility for S by U’s lights, out of a desire to live up to S’s trust. T will want to refrain from waiving because she will prefer that, if U comes to prefer T’s preferred option, U chooses that option, which is more likely if U thinks that he must do so on pain of exceeding his discretion. So whenever S and T have incompatible preferences concerning U’s decision, they will both disagree about whether S is to waive and about whether T is to waive, regardless of S’s preferences.

Thus I can no longer claim that it is the opposition of preferences concerning U’s action, and not their preferences concerning each other’s waiving choices, that robustly put S and T at odds, and that this is why they may be said to disagree about U’s action. Their opposition of preferences concerning each other’s waiving choices is robust with respect to changes in U’s preferences. However I think I can still claim that it is the opposition of preferences concerning U’s action that ‘really’ puts S and T at odds, and hence that it is natural that we should think of them as disagreeing about U’s actions.
Obviously, each only prefers that the other should yield because there is a chance that this will help to bring it about that U’s choices better satisfy her weighted desires—the constraints on their weighted desire-types rule out any other grounds they might have for their voiceable preferences concerning each other’s waiving choices. As before, it makes sense for the parties to *talk* as if they disagreed about U’s choices, rather than about each other’s waiving choices, because (a) it is their voiceable preferences concerning U’s choice that ‘really’ put them at odds, and (b) talking to each other about U’s choice is an economical way to simultaneously negotiate their two disagreements about each other’s waiving choices.
Bibliography


