The Constitution of Constitutivism

Karl Olof Leffler

Submitted in accordance with the requirements for the degree of Doctor of Philosophy.

The University of Leeds
School of Philosophy, Religion and History of Science

September, 2019

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Chapter 2, section 4, is based on material from:

- Leffler, O. 2016. The Foundations of Agency – and Ethics? *Philosophia*. 44(2), pp. 547-563.

Chapter 2, section 5, features some content (but does not reproduce any text) from:

- Leffler, O. 2014. *Rationalism with a Humean Face*. MA Dissertation, University of Gothenburg, SWE.

Chapter 3 is, with minor revisions, identical to:

- Leffler, O. 2019. New Shmagency Worries. *Journal of Ethics and Social Philosophy*. 15(2), pp. 121-145.

Chapter 6, section 1, and chapter 7, sections 2 and 3, feature material forthcoming in:

- Leffler, O. Forthcoming. Reasons Internalism, Cooperation, and Law. In: Garcia, M., Mellin, R., and Tuomela, R. eds. *Social Ontology, Normativity, and Law*. De Gruyter, GER [further information not yet available.]

The candidate is grateful to *Philosophia* and the *Journal of Ethics of Social Philosophy* for allowing him to republish material from Leffler (2016) and (2019), and to de Gruyter for allowing him to reproduce material from Leffler (forthcoming). He is also grateful to the University of Leeds for allowing him to reuse some ideas from Leffler (2014).

Acknowledgements

This dissertation has been long in the making. Two sets of events during my undergraduate years made me interested in the question 'Why be moral?'. The first was reading Immanuel Kant's *Groundwork to the Metaphysics of Morals* and Christine Korsgaard's *The Sources of Normativity* on train rides towards and from Lund, Sweden, where I spent part of my undergraduate days. Kant and Korsgaard made me ask the normative question: why be moral – or might morality be a mere chimera?

Later, at the Swedish Congress of Philosophy 2011 (*Filosofidagarna 2011*) at Gothenburg, I attended a talk by Emma Beckman on the error theory about morality. The Q&A session featured an exchange between her and her then-supervisor Gunnar Björnsson about the extent to which the idealization of an ideal agent in desire-based forms of reasons internalism could be expanded. I started to realize that there might be a way in which one could answer the normative question about morality by appealing to the ideal features of such agents. My MA dissertation, also advised by Gunnar, later took on that issue by discussing Michael Smith's attempts to extract moral reasons out of such theories. And even though my argument in this dissertation has developed a lot since then, I still fundamentally make use of that strategy to answer the normative question.

I want to thank everyone just mentioned for contributing in one way or another to the development of my thoughts. (And, even though I suppose I am now explicitly a Humean, that includes Kant.)

Moreover, the material here has been aired at many occasions since I started to think about it as a dissertation project. Audiences at Leeds have seen most of it, in one form or another. Audiences at Gothenburg and Linköping (*Filosofidagarna 2015*) have seen material from chapter 2. Audiences at the universities of York (*WRPPF*), Sheffield (*Understanding Value V*), and Humboldt-Universität zu Berlin have all seen versions of chapter 3. Audiences at LMU (*ECAP 9*) and Budapest (*Humeanisms*) have seen versions of chapter 4. Audiences at the universities of Kent (*Future of Normativity*) and Sheffield (*Understanding Value VII*) as well as UCD (*Irish Philosophical Society Annual Meeting 2018*) have seen versions of chapter 5. Yet another audience at Sheffield (*Understanding Value VIII*) has seen parts of chapter 6. And an unsuspecting audience at the University of Glasgow (*Social Ontology, Normativity and Philosophy of Law*) has seen parts of chapter 7. I want to thank everyone who has attended my presentations for listening and providing feedback.

There are, of course, individuals I have worked with more directly who also deserve credit. At Leeds, I especially want to thank my supervisors Pekka Väyrynen and Daniel Elstein, as well as my original supervisor Ulrike Heuer. I also want to thank the other members of CEM and everyone else in the School of PRHS for comments on various iterations of my arguments. In particular, I want to thank the postgraduate community, from which John Parry, Morten Højer Jensen, Will Gamester, Andrea White, Alison Toop, Adina Covaci, Gary Mullen, David Heering, Tadhg Ó Laoghaire, Pei-Lung Cheng, and Alexios Bréhier-Stamatiadis have provided particularly helpful ideas and feedback. So has Sakarias Bank with input from psychology.

Many philosophers from elsewhere should also be credited. When I was a visiting doctoral researcher at Humboldt-Universität zu Berlin in 2017, I was fortunate enough to meet and discuss aspects of this dissertation with Thomas Schmidt, Björn Petersson, Kathryn Lindeman, Jay Wallace, Razvan Sofroni, Berit Braun, Daniele Bruno, Francesca

Bunkenborg, and, most importantly, Janis Schaab and Benjamin Kiesewetter, who have given me especially helpful input. I want to thank everyone just mentioned – and the European Union for setting up the Erasmus+ scheme, which funded my stay and research activities in Berlin.

Beyond Leeds and Berlin, I am also indebted to the graduate community at Gothenburg during my MA days, James Lewis, William Hornett, Clemens Schlink, Neil Sinhababu, James Lenman, Luca Ferrero, Caroline Arruda, Neil Roughley, Herlinde Pauer-Studer, and Kirk Ludwig for comments, questions, and access to published or unpublished material. There are probably others I have forgotten too, and I apologize for that.

Last, but of course not least, some special mentions. When I applied for PhD positions, Caj Strandberg advised me that I should write about the potential dissertation topic that seemed the most interesting to me. I might not have written this dissertation if it had not been for that. I also want to thank my friends around the world, but I will not name names to make sure that I do not forget anyone. But I do want to mention Gabriel Nyberg's couch in Gothenburg, where a surprisingly large percentage of my better writing has been done, and my parents Klas and Marianne, who have been extremely supportive.

Abstract

Why be moral? According to constitutivism, there are features constitutive of agency, actual or ideal, the properties of which explain why moral norms are normative for us. I aim to investigate whether this idea is plausible.

I start off critically. After defining constitutivism and outlining its attractions and problems (chapter 1), I discuss the theories of various features of agency that are supposed to ground morality according to the leading constitutivists in the literature. I find these theories wanting. They are based on implausible assumptions about agency (chapter 2), and they fail to make sense of moral (and other) norms because the so-called shmagency objection, according to which we can shirk from our normative commitments by being 'shmagents' rather than agents, appears in new ways for them (chapter 3).

Then I get more constructive. I defend a two-tiered form of constitutivism. The first tier captures practical rationality, and the second tier captures reasons for action. Starting with the first tier, I defend a Humean theory of agency (chapter 4) and add a principle of instrumental rationality to it (chapter 5). Appendices A and B supplement these chapters with replies to criticisms of the Humean picture.

In chapters 6 and 7, I put this conception of agency to work to reach the second tier of the view. I start by defending a form of reasons internalism which treats practical reasons as grounded in the desires of ideal agents (chapter 6). Then I extend this theory to moral reasons, arguing – unexpectedly for a Humean – that we have universally prescriptive reasons to cooperate with other cooperative agents to satisfy our other respective desires (chapter 7). Hence, the constitutive features of ideal agency ground a morality of cooperation. I conclude by summarizing the case for constitutivism (chapter 8).

Table of Contents

0. Theses and Abbreviations

I use three conventions for abbreviating theses in this dissertation. These are:

(1) I use conventional abbreviations for many well-known philosophical theses.
Example: *HTM* stands for the Humean theory of motivation.

(2) When I introduce important theses of my own, for simplicity and readability, I capitalize their names.
Example: *PARADIGMATIC AGENCY* stands for what I take 'paradigmatic agency' to involve.

(3) When I modify theses of either kind, I add the modification in a lower bracket.
Example: *PARADIGMATIC AGENCY$_{IR}$* stands for *PARADIGMATIC AGENCY* when that is interpreted as including instrumental rationality.

Below, I list all the abbreviations in this dissertation as well as the most important theses picked out using conventions (1)-(3). I sometimes also denote theses that others have defended or various cases and *explananda* in similar ways, but these are handled very straightforwardly in the main text, so are not of much interest as far as the elucidation of my terminology is concerned. Hence, I only mention the abbreviations I use and my own important theses here.

(*A+*) An idealized agent with a psychology which explains an actual agent's reasons.

(*CENTRAL HUMEANISM*) A necessary feature of what makes one central class of events intentional actions is that a belief/desire-pair, suitably linked up, non-deviantly is part of the cause of the events in question.

(*CI*) The categorical imperative, in general.

(*CLOSE*) *A+*'s idealized desires explain *A*'s reasons only if *A+*'s desires and other psychological states range over circumstances that are similar enough to those *A* may be in.

(*DESIRES INTERNALISM*) For all *r*(*F,P,A,C*), *r(F,P,A,C)* is a reason relation holding between a fact *F* and a paradigmatic agent *P*'s action *A* in circumstances *C* iff (and because) *r*(*F,P,A,C*) holds in virtue of the desires that feature in *P*'s idealized psychology.

(*FOH*) The formula of humanity formulation of *CI* (roughly: 'act only so that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means').

(*FORMAL CONSTITUTIVISM*) An account *T* of some normative phenomenon *P* is constitutivist iff *T* entails that *P* is normative because *P* is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A** , where *C* constitutes something as an *A**.

(*FORMAL CONSTITUTIVISM$_P$*) An account *T* of some normative phenomenon P_P is constitutivist iff *T* entails that P_P is normative because P_P is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A**, where *C* constitutes something as an *A**.

(*FORMAL CONSTITUTIVISM$_{PM}$*) An account *T* of some normative phenomenon P_{PM} is constitutivist iff *T* entails that P_{PM} is normative because P_{PM} is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A**, where *C* constitutes something as an *A**.

(*FUL*) The formula of universal law formulation of *CI* (roughly: 'act only in a way such that your maxims could be made into universal law').

(*HI*) The hypothetical imperative, in general (roughly: 'taking means to one's ends is necessary for forming a will, on pain of irrationality').

(*HTM*) The Humean theory of motivation, in general (roughly: 'an action is an action in virtue of being non-deviantly caused by a belief/desire-pair').

(*IR*) If *A*'s end is to φ, and *A* believes that ψ is a necessary means to φ, and *A* does not take the means ψ, then *A* is instrumentally irrational.

(*IR$_{HUMEAN}$*) If *A* desires to φ, and *A* believes that ψ is the best means to φ, and *A* does not have an instrumental desire to ψ, then *A* is instrumentally irrational.

(*PARADIGMATIC AGENCY*) A person *P* is a paradigmatic agent, i.e. the kind of agent who is able to perform paradigmatic actions, only if *P* has a relevant set of beliefs and desires, and *P* is such an agent at least partially in virtue of having them.

(*PARADIGMATIC AGENCY$_{IR}$*) A person *P* is a paradigmatic agent, i.e. the kind of agent who is able to perform paradigmatic actions, only if *P* has a relevant set of beliefs and desires as well as *IR$_{HUMEAN}$*, and *P* is such an agent at least partially in virtue of having them.

(*PARTIAL CONSTITUTIVISM*) A form of constitutivism is a form of partial constitutivism iff the constitutive features of some aspect of agency properties of which explain normative phenomena are normatively justifiable (or desirable, required, etc.) to instantiate, rather than only descriptively necessary to instantiate for one to instantiate that aspect of agency.

(*ROBUST*) *A+* must have psychological dispositions and capacities that remain the same over minor changes in the circumstances she may inhabit or otherwise have desires for what to do in.

1. The What, Why and How of Constitutivism

Constitutivism is, roughly, the metanormative view that there are features constitutive of agency, actual or ideal, the properties of which explain why normative phenomena (such as reasons, values, or moral norms) are normative. Accordingly, constitutivism is supposed to show why these phenomena apply to us, and why they have normative force for us.

Constitutivism has many attractions. Not least, it is supposed to respond to various sceptical metanormative arguments, whether these are metaphysical or epistemological. Moreover, constitutivism is often considered to be able to solve various first-order normative problems. It might, for example, provide a rallying cry against normative disagreement, or help explain why agents are morally responsible even though they only may have reason to do what they can be motivated to do. So constitutivism has significant potential theoretical benefits.

In fact, the benefits of constitutivism need not just be (meta)normative. A constitutivist metanormative view may also be motivated as an upshot of our understanding of agency, for constitutivism may well be true just because it follows from a (good) explanation of agency. Accordingly, we can learn a lot about agency and various norms just by thinking about which norms, if any, might be constitutive of agency or otherwise explained by its constitution.

However, constitutivism also has all kinds of problems. For example, *the problem of substantive assumptions* says that the more substantive the assumptions a version of constitutivism is based on are, the easier it will be to deny one or some of them, hence refuting the view – but the weaker the assumptions some form of constitutivism is based on are, the less likely it is that it will be able to explain highfalutin normative phenomena like reasons or morality. *The problem of adequacy* is closely related to this problem; if an assumption fails, the version of constitutivism for which it fails is *ipso facto* inadequate. The *agency-shmagency problem* suggests that we can be shmagents, or not-quite-agents, and hence avoid the normative phenomena that supposedly are explained by what is constitutive of agency. And the *problem of bad action* suggests that constitutivism has a hard time explaining what it is to act badly because action is at least partially normatively constituted, so failing to follow the norm(s) that constitute(s) it seems to imply that one does not act.

These four problems structure this dissertation. I aim to develop them and show how standard forms of constitutivism suffer from them. But then I also develop a form of constitutivism that can solve them, as well as make good on the positive motivations for going constitutivist. This will not be enough for a conclusive defence of constitutivism; I will only show that it is tentatively justified and can be helpfully adapted in many metaethical frameworks. But, along the way, my argument will hopefully deepen our understanding of how a good version of constitutivism should look, as well as of the metanormative aspects of agency and associated phenomena more generally.

My methodology takes inspiration from the Canberra plan. It does so in the broad sense that I often work by starting with some data points, platitudes, conceptual truths, or other, perhaps more theoretical, features of some phenomenon, and then try to see what, if anything, best captures them.[1,2] Often, this will have me trying to make the most minimal assumptions possible, strengthening the case for constitutivism by arguing from parsimony, but sometimes I will instead emphasize other theoretical virtues or benefits of the view. This way of working should be uncontroversial; from an abstract enough perspective, it is compatible with most standard philosophical methodologies, including the method of reflective equilibrium applied in general.

To show how constitutivism is theoretically virtuous or beneficial, I shall start here in chapter 1 by clarifying what constitutivism is in greater detail, along with its benefits and potential problems. Then I get destructive. In chapter 2, I present what I take to be the four leading versions of constitutivism about moral norms in the literature, viz. Korsgaard's, Velleman's, Katsafanas' and Smith's. I argue that they suffer from the problem of substantive assumptions – and that their substantive assumptions about agency are, in various ways, inadequate.

In chapter 3, I present two new shmagency worries. First, versions of constitutivism that take the constitutive features agency that explain normativity to be descriptively instantiated cannot explain the reasons of sophisticated shmagents. Second, versions of constitutivism that take the features of agency to be normatively required of agents leave normative phenomena like reasons underdetermined. This means that all the leading forms of moral constitutivism suffer from some version of the shmagency objection.

[1] In particular, I do this in the second half of the dissertation when I discuss agency, rationality, reasons, and morality.

[2] However, I will often lack the space to discuss competing or complementary views in depth. In such cases, the reader will have to make do with the case I make for the views I defend – or read the dissertation hypothetically.

But then I turn constructive. Fundamentally, I aim to defend a two-tiered version of constitutivism. The first tier involves defending a principle of instrumental rationality as a constitutive feature of Humean agents – as well as its normativity. With the second tier, I shall explain reasons for action by arguing that they are grounded in the desires of idealized agents, where these are understood as in the first tier.

I defend the first tier in chapters 4 and 5. In chapter 4, I defend a Humean conception of agency, arguing that because many central (or 'paradigmatic') cases of action plausibly are actions in virtue of being caused by belief/desire-pairs, paradigmatic *agents* are plausibly constituted by (at least) the mental states that feature in such pairs. This chapter is supplemented by appendix A, where I reply to objections to the Humean theory of motivation.

Then, in chapter 5, I argue that a principle of instrumental rationality supplements the beliefs and desires of paradigmatic agents. I also argue that this form of instrumental rationality has normative force, and, in response to the problem of bad action, that one should think of action *disjunctively*. This chapter is supplemented by appendix B, where I respond to alternative accounts of rationality.

I turn to the second tier of my view in chapter 6. In spite of decades of criticism, there I argue that a form of desire-based practical reasons internalism, according to which reasons are reasons because they are grounded in agents' idealized desires, explain a host of features one may want to explain about reasons. In chapter 7, I extend that explanation to moral reasons. I argue that, as a matter of being ideal, fully idealized agents must desire to cooperate with other cooperative agents to satisfy their other respective (non-antisocial) desires. These desires ground a morality of cooperation. Hence, perhaps unexpectedly, my Humean constitutivism can explain at least some universally prescriptive moral norms. Finally, I conclude in chapter 8 by wrapping up the case for constitutivism.

But most of that will come later. Again, in this chapter, I shall introduce what constitutivism is and outline its benefits and potential problems in greater depth. I start in section 1 by discussing some standard definitions of constitutivism. In section 2, I instead propose my own definition, explain it, and then delineate the specific version of constitutivism I am after discussing; namely, constitutivism about moral norms. In section 3, I provide some reasons for thinking that constitutivism matters theoretically. In section 4, I introduce the main objections to constitutivism, setting the stage for the substantive chapters to follow. I conclude in section 5.

(1) What is Constitutivism?

Again, roughly, constitutivism is the metanormative view according to which there are features constitutive of agency, actual or ideal, the properties of which explain why normative phenomena are normative. What does that mean?

By 'features constitutive of agency', I mean the features of agency (or similar phenomena, e.g. action) that constitute agency (or the other phenomena) as the kind of thing it is (or they are). And by 'normative phenomena' (or 'norms'), I mean phenomena that we ordinarily take to have normative force for us, such as reasons, rational requirements, values, or moral norms. Hence, the core constitutivist idea is that there are certain features of agency (or some related phenomenon or phenomena) which have properties that explain why normative phenomena are normative. But just *what* is constituted such that it explains norms, or *which* normative phenomena are normative in this way, varies between different theories.

To make more sense of constitutivism, however, we need a more stringent characterization of it. Unfortunately, constitutivism is often discussed inconsistently or metaphorically. Paul Katsafanas, for example, says that constitutivism 'generates' reasons from constitutive features (Katsafanas, 2013, p. 1; 2018, p. 369). But he also says that it 'grounds' normativity in them (2013, p. 4; 2018, p. 367), that normative claims 'apply' to us in virtue of being agents, that features of action 'yield' normative standards of assessment for action, and that constitutivism 'justifies' normative claims because agents are committed to them (2018, p. 367). In the same anthology as Katsafanas (2018), Southwood (2018*a*, p. 345; cf. Smith, 2017) says that constitutivism 'constitutively explains' reasons. And Smith (2015, p. 187) says that constitutivists attempts to 'derive' reasons from action and agency – while discussing Wiland (2012, p. 117), who says that constitutivists attempt to 'extract' norms from constitutive features.

Inconsistency or metaphors will not do; they muddy the waters. Nor will it do to let writers self-identify as constitutivists, for labels can be misused or used inconsistently. And, importantly, constitutivism should not be run together with metaethical *constructivism* – constitutivists need not be constructivists. No one thinks that constitutivists about epistemic reasons, who hold that reasons are truth-conducive considerations because truth is the constitutive aim of belief, are constructivists about epistemic reasons. Moreover, the nature of constructivism is itself controversial, and hence not obviously helpful (cf. Bagnoli, 2013; Lenman and Shemmer, 2012; Southwood, 2018*a*; 2018*b*; Street, 2010).

So how should we understand constitutivism? Despite their at times less than helpful terminology, Katsafanas and Smith have also presented the two most exhaustive characterizations of constitutivism in the literature. Can they help? I shall start by discussing Katsafanas' view. He defines constitutivism by appealing to a schema:

> (*CONSTITUTIVE AIM*) Let *A* be a type of attitude or event. Let *G* be a goal. *A* constitutively aims at *G* iff (*i*) each token of *A* aims at *G*, and (*ii*) aiming at *G* is part of what constitutes an attitude or event as a token of *A*. (Katsafanas, 2013, p. 39)

A constitutive aim, in this case, means that there is a certain goal that all token instances of an activity – where the activity itself is made up of having various attitudes or events – share, and the aim is constitutive of what that activity is. Since constitutive features are necessary features, the tokens of that activity must have the aim, or else they do not count as partaking in the activity. The standard example of an activity with such an aim is chess. To play chess, one must follow its rules and aim at winning (or at least drawing), or else one is not really playing it (Katsafanas, 2013, pp. 1-2; 38).

Seemingly construing actions (or *A*) as events, Katsafanas thinks actions share this structure. Even beyond being suitably related to agents' individual aims, there are certain things events must be to count as actions. Agents have individual aims that they can try to reach – to catch the game, tie their shoes, eat a bagel; call these first-order aims – but because actions must meet extra conditions beyond the agents' aims to count as actions, there are also higher-order aims internal to the nature of action that hold for agents when they aim at their first-order aims. Because action in general has higher-order aims, anyone who acts *also*, implicitly, aims at the higher-order aims, even though agents need not aim at these higher-order aims intentionally or knowingly.

A good example of a theory which has this structure is J. David Velleman's form of constitutivism. Our first-order aims may indeed be to catch the game, tie our shoes, or eat a bagel, but self-knowledge is a higher-order aim that all actions also aim at (cf. Velleman, 2000; 2007*a*; 2007*b*; 2009; cf. chapter 2, section 3 below). For Velleman, an intention is a kind of belief that itself provides us with knowledge about what we do when we act, and hence knowing ourselves is a higher-order aim of all actions that feature intentions.

Katsafanas thinks a further claim follows from *CONSTITUTIVE AIM*:

> (*SUCCESS*) If *X* aims at *G*, then *G* is a standard of success for *A*. (Katsafanas, 2013, p. 39)[3]

SUCCESS shows why aims have normative import. The normativity here is one of reasons; Katsafanas explicates *SUCCESS* further by claiming that if *X* aims at *G*, *X ipso facto* has reason(s) to live up to it. (What relation holds between the aim and the reason is unclear, but just assume the claim for now.) Since action has constitutive aims, then, *SUCCESS* is inescapable insofar as we act. Hence, insofar as we act, we are subject to the constitutive aims of actions as success conditions for the actions – and because they are such conditions, they provide us with reasons. So, for Katsafanas, a constitutivist account of some normative phenomenon is anything that satisfies *CONSTITUTIVE AIM*, and hence also grounds the normative *SUCCESS*, interpreted in terms of reasons.

However, quite generally, a good characterization of a theory needs to be precise by clearly capturing what it is supposed to capture, and it needs to be extensionally adequate by capturing the theory as it is discussed in the literature. Katsafanas' characterization of constitutivism lacks both these features. I have three worries about the precision of Katsafanas' proposal as well as one major worry about its extensional adequacy.

First, Katsafanas is unclear about whether he is talking about grounding normativity in what is constitutive of action or agency. He oscillates between the two terms (cf. e.g. Katsafanas, 2013, p. 1 vs. p. 46; Katsafanas, 2018, p. 367). On one reading, where an agent is anyone who acts whatsoever, and anyone who is an agent *ipso facto* acts, the ambiguous language could be forgiven. Here, the formulations would refer to the same thing.

But this reading is misleading insofar as we are discussing constitutivism. A view where several aspects of action and agency come apart is logically possible; an agent can have more features than those that are involved in initiating some particular kind(s) of action. To precede the discussion in chapters 4 and 5: if most actions require belief/desire-pairs, as per standard Humean action theory, but some actions can be performed out of desires without any accompanying beliefs, it might be that an agent

[3] However, later in the same book, Katsafanas qualifies *SUCCESS*, claiming that his version of constitutivism only requires that if an agent aims at some goal, and endorses this aim, achieving that goal is a standard of success for the action (Katsafanas, 2013, pp. 208-209).

necessarily has beliefs and desires – *and* can perform some non-Humean actions without beliefs along with the standard Humean ones. Hence, we should not think that action and agency line up seamlessly so that one can jump unproblematically between talking about one and the other. Katsafanas' interpretation of *A*, then, seems underdeveloped.

A second issue emerges when Katsafanas talks about how constitutivism 'generates', 'grounds', 'yields', or 'justifies' normativity, or shows how it 'applies' to agents. I presume that if a norm 'applies' to someone, that means that she is subject to it. But no one in the literature does, to my knowledge, discusses 'generation' or 'yielding' relations at much length, so I assume these terms are metaphors for some kind of normative explanation. But which one? Grounding? If so, does Katsafanas intend to talk about some ambitious grounding relation when he talks about grounding (and if so which one?), or does he talk more loosely? And how does *justification* come into the picture?

Katsafanas does not say, but regardless of what he thinks, constitutivists should at the very least not be committed to grounding in any strong sense of the word. Normative grounding is controversial (cf. e.g. Väyrynen, 2013). Instead, constitutivists should be open-minded about which relation they appeal to here.

Third, Katsafanas does not attempt to unpack his terminology when he discusses the *explananda* of constitutivist views. But what kind of normative phenomena do constitutivists attempt to 'generate' (or 'ground', etc.)? Are we talking about reasons, rational requirements, moral norms, or something else? Interpreting *SUCCESS* in terms of reasons seems very presuming. Pre-theoretically, it should be possible to be a constitutivist about many different normative phenomena.

Apart from these worries about the precision of Katsafanas' view, I have a major worry about the extensional adequacy of his schema. Quite generally, constitutivist explanations can come from different kinds of constitutions, so it seems like there is no need to define constitutivism in terms of the constitutive *aims* of actions. There are, at least, three problems with doing so.

First, Katsafanas himself has conceded that we may understand what is constitutive of action in terms of 'principles', meaning that there might be a deontic rule such as the categorical imperative (*CI*) which is constitutive of action, but that we do not necessarily *aim* at living up to. He is not fully clear about what the aim/principle distinction involves, but the difference seems to be that aims are set by some psychological forces (e.g. desires or drives) or values, whereas principles are rules that have little to do with such factors (Katsafanas, 2013, ch. 4; cf. Katsafanas, 2018). If so, there can be at least two forms of constitutivism: aim-based or principle-based.

Second, one need not interpret actions as events with constitutive aims. It is unclear whether actions are events in general; many writers deny this. Indeed, Katsafanas *himself* often writes as if he takes actions to be processes (cf. Katsafanas, 2013, ch. 6; cf. chapter 2, section 4 below). More specifically, he believes that drive-motivated actions are processes, and all actions are drive-motivated. So perhaps actions are *processes* with aims or principles rather than events. But it is not obvious whether these can be reduced to events (cf. Steward, 2012).

Third, there might be structural features of agency (or similar phenomena) such that what emerges from them counts as normative, quite independently of whether these features have anything to do with what we aim at or the reasons we might acquire by aiming at something. By a 'structural feature', I mean something that everyone counting as something *S* non-accidentally possesses in virtue of counting as *S*, and either is what constitutes them as *S* or is required by what constitutes them as *S*. For example, on Michael Smith's view (2011; 2012*a*; 2013; 2015; 2017; 2018), ideal agents are partially constituted by having certain desires – but most agents do not *aim* at having those desires. Rather, ideal agents are ideal agents in part because they have them.

Katsafanas' view seems problematic in many ways, then. Does Smith do better? Inspired by Thomson (2008), he holds:

> (*GOODNESS-FIXING KIND*) An account of some normative fact is constitutivist when the account explains that normative fact in terms of some goodness-fixing kind. (Smith, 2017, p. 373)

GOODNESS-FIXING KIND needs explication. The core idea is that a goodness-fixing kind is a kind that itself sets out the features that something has to have to be a good member of it. The standard example is a toaster: a toaster is functionally supposed to toast bread; if it does that well, it is good *qua* toaster (and an exemplar of its kind); if it does not, it is bad *qua* toaster. This means that one can rank all instances of toasterhood by the extent to which they live up to the properties of toaster exemplars, generating an ordering of better or worse toasters in virtue of their degree of deviation from the exemplar.

There are questions to ask about how kinds might be able to do that. One possibility is that some sort of functionalism is implicit in our concept of some kind (e.g.

the concept of a toaster), and then we can rank instances of toasters relative to the conceptual function. Another possibility is the Aristotelian idea that some (kinds of) objects metaphysically are functionally organized entities. Whatever we think about artefacts like toasters, perhaps vision is the function of eyes, and eyes function better the better they let someone see. To make use of *GOODNESS-FIXING KIND*, one could then attempt to argue that agency, action, or something like them have functions in at least one of these ways – and we should explain norms in terms of these functions.

GOODNESS-FIXING KIND might seem to be better than *CONSTITUTIVE AIM* when it comes to the objections I presented above. When it comes to the precision of the views, first, both actions and agents could plausibly be goodness-fixing kinds. Second, it is clearer how the normative explanation would look here than on Katsafanas' 'generation' vs. 'grounding' formulations – it would be based on functional goodness. And third, *GOODNESS-FIXING KIND* seems like a promising theoretical candidate for explaining many different normative phenomena. Moreover, the view is not limited to taking constitutive aims to be properties of attitudes or events in the way Katsafanas limits his view, and hence seems extensionally better than *CONSTITUTIVE AIM*. While attitudes or events could have functional aims, perhaps there are also other (conceptual or metaphysical) kinds of entities (e.g. agency itself) that have functional aims. If so, it is possible to formulate versions of constitutivism based on those kinds of entities.

However, *GOODNESS-FIXING KIND* suffers from two extensional problems. First, it fails to include key constitutivists in the literature under its umbrella, hence seeming too exclusive. Smith (2017) is happy to rule out leading constitutivists like Velleman and Katsafanas from his definition because – he thinks – they appeal to constitutive aims rather than goodness-fixing kinds. Smith thinks that agents either need not have constitutive aims or that talk of such aims seems theoretically revisionary. However, given the importance aims have in several well-known constitutivist frameworks, this move seems objectionably limiting.

Second, the view also seems too inclusive. Smith's formulation rules in non-naturalist quietist realist T.M. Scanlon (1998; 2014) as a constitutivist about reasons for action (Smith, 2017). Smith's idea is based on taking Scanlon to explain reasons for action by appealing to (the goodness-fixing kind) 'deliberator'. Something is a reason for action, on this view, if a (good) deliberator has an intention to perform it, even though there are irreducible reasons for *attitudes* that this deliberator must be sensitive to when forming those intentions. Scanlon is, however, not regarded as a constitutivist by anyone else in the literature. I think this is correct, because on Scanlon's view, at least Smith understands it, all the normativity-explaining work behind our reasons for action is done by the irreducible reasons for attitudes. So *GOODNESS-FIXING KIND* rules in too much.

(2) What (Moral) Constitutivism Is

Here is a schema that improves on Katsafanas' and Smith's definitions:

> (*FORMAL CONSTITUTIVISM*) An account *T* of some normative phenomenon *P* is constitutivist iff *T* entails that *P* is normative because *P* is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A** , where *C* constitutes something as an *A**.

Clarification is needed. By *T*, I mean a theory that aims to explain the normativity of *P*. A 'normative phenomenon' (or 'norm', for short) *P* could be anything conventionally understood as normative, such as reasons, rational requirements, moral norms, or goodness-for. By '*P* is normative', I mean why *P* holds for or applies to an agent and has normative force for her, where 'holding for or applying to an agent' indicates that an agent is subject to the norm, and 'has normative force for her' means what it says.[4] Some philosophers use the term 'normativity' only for the latter property, and I shall sometimes use it that way too, in line with the conventions in the literature. But I shall only do so when the context makes it clear that that is what I have in mind; most of the time, I mean to use the term to indicate that a norm applies *and* has force.

Furthermore, *C* is either constitutive aims or principles, as per Katsafanas' definition, or some other constitutive structural feature or features of agency, understood as in the last section. And the aspect of agency denoted by *A** could be anything conventionally associated with agency, such as action, agency itself, propositional attitudes, or selfhood. However, for simplicity and readability, I sometimes lump these aspects of agency together under the umbrella term 'agency'.

Some features of this definition need more explication still. First, it is important that *P* is normative in virtue of some property or properties of the constitutive features *C*. *P* need not be normative in virtue of the constitutive feature or features themselves.

[4] This leaves open questions: it is unclear *who P* is supposed to be normative for, or whether all allegedly normative phenomena *are* normative in this sense. Constitutivists may legitimately disagree on these issues.

The constitutive features need only serve to transmit normativity from some other source; indeed, most constitutivists seem to think that some aspect of agency is (independently) valuable or inescapable, and that *that* is what explains why *P* is normative. This is so even though value or inescapability need not be constitutive of agency themselves (Ferrero, 2018; 2019; cf. chapter 3; 5 below).

Second, the constitutive features *C* properties of which explain the normativity of *P* might themselves be some aspect or aspects of agency, and normative for that reason. However, *C* could also be normative in virtue of some *other* aspect of agency. This means that constitutivist views can have multiple tiers, where different normative phenomena may have different roles according to different forms of constitutivism. Michael Smith's view is a good example here (cf. chapter 2, section 5; chapter 3, section 5; 6). He thinks reasons are explained by the desires of ideal agents, but ideal agency is normative because agency is a goodness-fixing kind – which in itself needs independent explanation. This generates a two-tiered view.

Moreover, constitutivism need not be wedded to constitutive explanations of normativity. The relation that holds between *P* and constitution is different; constitutivism says that *P* is normative in virtue of some property or properties of the constitutive feature(s) *C* of *A**. There is indeed a constitution relation here, but it holds between *A** and *C*. This relation need not be the same relation as the one that makes *P* normative.

Instead, maybe some other explanatory relation holds between some properties of *C* of *A** and *P*, and hence shows why *P* is normative. Here, I interpret 'explanation' loosely: *F* explains *G* if *G* is so because *F* is so (cf. Broome, 2013, pp. 48-49). This allows for many possible explanatory relations: *G* could be constituted by *F*, be grounded (or 'sourced') in *F*, be identical with *F*, etc. This is helpful because different constitutivists handle their fundamental norms in different ways, and here *P* can be normative in virtue of many different relations.[5] It is, in fact, also possible that a relation like grounding might *back* explanations of the normativity of *P* rather than be explanatory itself, though I shall proceed to use the explanation terminology here for simplicity.

Finally, I do not intend to say that all constitutivists believe that some aspect of agency can capture all normative phenomena *P*. Different constitutivists attempt to explain different aspects of the normative sphere. For example, Sharon Street (2008)

[5] One might even construe the constitution relation in a way where it explains non-natural normative facts (Shafer-Landau, 2003, ch. 3). If that is what 'constitution' requires, constitutivism seems, surprisingly, to be compatible with non-naturalism.

believes that practical reasons (*P*) can be explained in terms of the right kind of value judgments, which are a kind of propositional attitudes (*A**) constituted by being made from a practical point of view (*C*). On the other hand, Christine Korsgaard (2009) believes that *CI* (*P*) is constitutive of human agents (*A**), as human agents inescapably universalize their maxims by it in a way which enables them to stand behind their actions (*C*). Both are constitutivists.

FORMAL CONSTITUTIVISM can solve all the problems that I raised for Katsafanas' and Smith's characterizations of constitutivism. Regarding the problems for Katsafanas, first, *FORMAL CONSTITUTIVISM* allows for more variations regarding *A**, *inter alia* disentangling action and agency. Second, it allows for several possible relations between constitutive features and normative phenomena. Third, it can feature several kinds of normative phenomena. And fourth, *C* can be other constitutive features than aims.

Moreover, regarding the problems for Smith, *FORMAL CONSTITUTIVISM* seems more extensionally accurate than *GOODNESS-FIXING KIND.* As I will show in chapter 2, it counts leading constitutivists as constitutivists. But it does not include Scanlon, for it is not the properties of an agent who is sensitive to reasons for attitudes which explain the normativity of reasons for actions on his view, but rather the force of the reasons for attitudes themselves.

So *FORMAL CONSTITUTIVISM* seems promising. But it can be further narrowed down to explain different types of normative phenomena. For example, we may be interested in:

> *(FORMAL CONSTITUTIVISM*$_P$*)* An account *T* of some normative phenomenon P_P is constitutivist iff *T* entails that P_P is normative because P_P is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A**, where *C* constitutes something as an *A**.

By P_P, I mean normative phenomena that are conventionally recognized as distinctively practical. By this, I mean phenomena associated with action, not (ordinary) beliefs or semantic meaning. Typical examples are goodness-for, practical rationality, morality, or practical reasons.

We can narrow *FORMAL CONSTITUTIVISM*$_P$ down even further to reach morality:

> (*FORMAL CONSTITUTIVISM$_{PM}$*) An account *T* of some normative phenomenon *P$_{PM}$* is constitutivist iff *T* entails that *P$_{PM}$* is normative because *P$_{PM}$* is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A**, where *C* constitutes something as an *A**.

By *P$_{PM}$*, I mean a *moral* practical normative phenomenon. Hence, I distinguish between constitutivist views that are or entail moral norms and other practical constitutivist views. I shall primarily concern myself with moral forms of practical constitutivism like this in the rest of this dissertation, but I shall occasionally also touch on other forms of *FORMAL CONSTITUTIVISM$_{P}$*, such as when I discuss rationality or reasons for action.

What differentiates moral constitutivism from other forms of (practical) constitutivism? Several suggestions about how to distinguish moral norms from non-moral ones exist in the literature. Some even deny that morality is categorically distinct from other kinds of normative phenomena, e.g. practical reasons in general. I do not have the space to discuss all these suggestions in depth here. Yet, following Smith (1994, ch. 5), I suspect that there are two properties that are the strongest marks of the moral – universal prescriptivity and conventional recognizability.[6]

Most philosophers agree that moral norms are prescriptive in the sense that they have some sort of normative 'force', 'oomph', or 'to-be-done-ness' built into them, though how that property should be understood is highly controversial. I call this elusive property 'prescriptivity' or sometimes 'normativity' (which should not be confused with 'normativity' in the sense in which it features in *FORMAL CONSTITUTIVISM* – cf. the explication above), but the reader will have to wait until chapter 5 until I characterize it in much depth. Nevertheless, it seems highly plausible to me that moral norms have a kind of prescriptivity, whatever it is, that holds *for all*, or at least all minimally sophisticated, agents. If you and I both witness a child drowning in a pond, and we are such that we can save it without any significant cost, it is wrong for both of us not to try

[6] Based on a comprehensive literature review, Forcehimes and Semrau (2018) list four potential moral/non-moral distinctions: (*i*) moral reasons are not merely social, but have stronger force than that, (*ii*) moral reasons do not depend on individual commitments (but are categorical), (*iii*) moral reasons are responsibility-implying, and (*iv*) moral reasons are altruistic. Here, (*i*) and (*ii*) look like ways to try to spell out the intuition of universal prescriptivity, whereas (*iii*) and (*iv*) are ways to spell out conventional recognizability.

to do so. I shall call this kind of prescriptivity 'universal prescriptivity' and try to formulate a form of constitutivism that can capture it.[7]

Admittedly, some deny that moral properties have normative force or that the normative force of morality is weaker than being universally prescriptive. For example, Dorsey (2016; cf. Boyd, 1988; Brink, 1989) is a recent sustained critique of moral rationalism, taking that to be the view that morality necessarily gives everyone reasons. Similarly, on Street's (2008; 2012) constitutivist view, moral reasons fundamentally depend on value judgements that at best are contingently universal, but mostly likely are not. I think these philosophers have got the data wrong; morality, as we understand it, involves a claim to universal prescriptivity, so insofar as these philosophers think that morality has weaker normative force than that, I think they change the topic.

However, those who hold that morality has some weaker form of prescriptivity, or even none at all, may want to read the rest of this dissertation hypothetically, as an attempt to answer the question: can we explain moral normativity in a way that shows how moral norms are universally prescriptive, and hence show how morality has that type of normativity, whether or not that is what our ordinary discourse seems to involve?

The second property that characterizes morality as I understand it is conventional moral recognizability. A constitutivist view about moral norms should explain how conventionally recognizable moral norms emerge based on the constitutive features of some aspect or aspects of agency – whether they do so directly, if some constitutivist aspect of agency *is* recognizably moral, or less directly, so that moral norms gain support from some constitutivist aspect or aspects of agency. *CI* is a good candidate for the former kind of explanation, whereas Velleman's self-knowledge norm is a generates a version of the latter type of explanation (cf. chapter 2).

But what is 'conventional recognizability'? I leave that very open. The only thing that matters is that a norm can be picked out as moral according to some moral convention, which in turn may be either a folk convention or something more theoretically informed. For perhaps there are features of moral conventions that philosophers have yet to pick up on, but as long as there is *some* established convention whereby some norm plausibly can count as moral, that should be enough to count it as moral. This is because I do not want to be too restrictive about what counts as moral – I do not want to take much of a stand on contentious first-order moral issues by ruling

[7] It is no surprise that I talk about 'universal prescriptivity' here. My 'universal' adjective captures what Hare (1981, ch. 1) means by universalizability. I am not committed to Hare's interpretation of the other properties he thinks are constitutive of moral judgements, i.e. prescriptivity and overridingness, however.

various views out pre-theoretically. Writers often seem tempted to do that. For example, Wallace (2019; cf. Mayr, 2019) thinks that it is a necessary feature of moral obligations that we owe them to others *as equals*. But this seems like a substantive and possibly deniable first-order judgement to me.

Indeed, I am pre-theoretically agnostic about what first-order morality amounts to, so it is not admissible for me to bake conclusions about what it involves into my framework pre-theoretically. It is therefore important that the intuitions delimiting conventional recognizability are not first-order moral intuitions, but rather conceptual or linguistic intuitions about how we (or people of some tradition) usually think or talk. And even then, they are more of a guiding light for what we might be interested in explaining than *desiderata* that generate reasons to dismiss a view that cannot explain them. If something else than morality as it is conventionally understood would happen to be universally prescriptive in virtue of properties of features constitutive of some aspect of agency, then maybe we should stick with that normative phenomenon, as opposed to – or maybe together with – morality in our actual practices.[8]

(3) Why Constitutivism?

Constitutivism has gained a lot of traction during the last decades. But why does it – and especially *FORMAL CONSTITUTIVISM$_{PM}$* – matter? In this section, I present the main reasons for thinking that it does. In chapter 8, I will attempt to show how the view I shall develop in the coming chapters makes good on them.

The first reason for why many philosophers are attracted to moral constitutivism is metaethical. Constitutivism promises to explain the normativity of moral norms in an attractive way, hence solving various sceptical problems and vindicating morality. Many constitutivists have proposed versions of *FORMAL CONSTITUTIVISM$_{PM}$* when they have attempted to explain universally prescriptive moral norms either in terms of less controversial normative premises than moral ones, or in terms of completely non-normative premises (cf. chapter 2).

Such explanations would be able to capture the normativity of moral norms even in the light of widespread moral scepticism stemming from various metaethical considerations. Which forms of scepticism and considerations are these?

[8] For the same reason, it does not matter to me, theoretically speaking, if I end up being partially morally revisionary. This is very helpful, as my explanation of morality in chapter 7 might not be able to capture everything about it.

Many arguments have been presented against standard metaethical theories that purport to vindicate morality – in particular against various forms of moral realism. First, there are metaphysical arguments, not least Mackie-inspired arguments from queerness, according to which moral discourse appears committed to moral properties if it is to be vindicated, but those properties are too queer to exist (Mackie, 1977; Joyce, 2001; Olson, 2014). It is not obvious how this error-theoretical argument should be interpreted, but most significantly, many error theorists tend to hold that moral norms must have the right kind of universal prescriptivity – they must be have normative force, and have it for all – but that there are no properties which have that kind of force. Upshot: there are no moral properties.

Another, related, worry is that normative phenomena, moral or other, may be wholly irreducible to something descriptive. If so, it might seem like *any* kind of normativity is too queer to exist. I will not be discussing this issue in much depth, but when I sum up my case for constitutivism in chapter 8, I shall endeavour to show that the view I have defended provides the right kind of solution to this problem too.

It is sometimes claimed that other metaphysical aspects of moral properties than those just mentioned make them queer too. These tend to involve the supervenience of the moral on non-moral or descriptive properties, or the alleged ability of moral properties to be intrinsically motivational (cf. Olson, 2014, ch. 6-7). I am less fazed by these other arguments; there are many theoretical options about moral supervenience (cf. McPherson, 2015), and I am not sure about whether motivational internalism even is a datum that should be explained in the first place. It is possibly true, but also possibly not, and the issue remains extremely controversial (cf. e.g. Björklund *et al.* 2015; Brink, 1989, ch. 3; Korsgaard, 1986; Smith, 1994, ch. 3). Hence, I shall not discuss these potential aspects of moral properties further.

More importantly, however, there are also epistemological arguments for moral scepticism. One of these features in Mackie's original argument from queerness – he thinks that moral knowledge would have to come from a queer form of intellectual intuition (Mackie, 1977, ch. 1). It is unclear whether *that* particular challenge is very worrying anymore. Intuitionism has seen a resurgence (cf. e.g. Shafer-Landau, 2003; Huemer, 2005; Enoch, 2011*a*), and there are other theoretical options that might capture the foundations of our moral knowledge. For example, it is unclear why one could not go with process reliabilism (Goldman, 1979), some kind of virtue epistemology (Sosa, 2007; 2009), or even a knowledge-first view instead (Williamson, 2001).

However, the dominant theory of epistemic justification in ethics is not foundationalist, but a coherentist interpretation of the method of reflective equilibrium. Put very briefly, according to this method, the way in which we ought to go about justifying our beliefs is by making them coherent with each other in the light of all potentially relevant justificatory factors, adjusting either our particular judgements or the principles underlying them. According to the coherentist interpretation of the method, any kind of beliefs can justify moral beliefs iff they are coherent with our other judgements (cf. Daniels, 1996; Rawls, 1971).

But both the method and coherentism about justification may be questioned: coherentism suffers from a plethora of well-known problems. For some examples, there is the garbage-in/garbage-out worry that a lack of good initial judgements will yield poor coherent belief sets (Smith, 2018), the general worry that coherence is not truth-conducive (Olsson, 2005), and the worry that there are several different possible coherent equilibria, making it unclear why one should accept any particular one (Tersman, 1992).

Even worse, versions of these problems appear regardless of which epistemology one uses to interpret the method of reflective equilibrium. Whether or not it is coherence that ultimately confers justification on beliefs, the state of reflective equilibrium at the end of inquiry reached by the application of the method of reflective equilibrium is liable to all these problems, *whether or not* one gives the method a coherentist interpretation. Poor initial judgements may always yield poor conclusions, coherence need not be truth-conducive, and it is still unclear which equilibrium one should go with.

Moreover, there are also other famous challenges in moral epistemology. First, many doubt the reliability of moral beliefs or intuitions in the face of disagreement, both in virtue of how widespread disagreement is over time and space, and because it seems remarkably persistent (cf. e.g. Katsafanas, 2013, ch. 1; Tersman, 1992). Moral disagreement often seems even deeper, more pervasive, and more theoretically challenging than disagreement on non-moral matters.

Second, it is not obvious how moral beliefs or intuitions are connected with an external moral reality (assuming, for now, that one exists) or whether that connection in fact at best may be some sort of lucky cosmic coincidence when there is no clear theory about what the connection amounts to (Street, 2006; Enoch, 2011*a*; cf. Bedke, 2009).[9]

[9] I follow Enoch (2011*a*, ch. 7) in treating Street's (2006) evolutionary challenge as a version of this cosmic coincidence objection. Readers who disagree may think of Street's challenge in other ways, but it is never a challenge for constitutivism – the properties of constitutive features of various aspects of agency that explain norms according to constitutivists are plausibly evolved.

And since it is widely held that knowledge cannot depend on lucky coincidences, even if our beliefs are true, such a coincidence would undermine our knowledge.

Hence, it would be helpful if constitutivism could give us *some* more epistemic purchase. I shall attempt to show its benefits on three fronts: first, how it can provide an anchoring point for moral reflection and therefore help us solve the problems I mentioned for the method of reflective equilibrium in general. Second, how it can contribute to solving problems of moral disagreement. And third, how it shows how moral norms are explained by constitutive features of agency that we can access, hence handling the cosmic coincidence worry.

Metaphysical and epistemological challenges aside, one may also think of even further sceptical worries about morality. For example, if moral properties have no causal impact on the world, and one wants to accept a causal theory of reference, it seems impossible to refer to them. This paves the way for semantic scepticism (cf. Benacerraf, 1973). But whether constitutivism can help with such extra arguments – constitutivism mostly has metaphysical and epistemic import, and it is therefore compatible with most semantic views – is a question best left for elsewhere. Hence, for now, I shall focus on the metaphysical and epistemic sceptical challenges.

However, the anti-sceptical motivation for constitutivism might seem to beg the question. Why would one want to vindicate moral norms if it seems like there are no objective ones due to these arguments for scepticism? Assuming that morality is valid enough to deserve vindication might look unwarranted, given their strength.

One answer is that the sceptical arguments can be viewed as challenging our pre-theoretical understanding of moral discourse. Even philosophers who, in the final analysis, deny the appearances tend to concede that we start off believing that moral discourse involves certain typical moral judgements, as well as some more abstract metaphysical and epistemic properties (e.g. that morality seems to be universally prescriptive, and that we tend to think that we know some moral truths). But if moral nihilism then looms, one alternative theoretical route is to look for constitutivist explanations of morality.

But are those explanations any good? That has yet to be established. The potential payoff constitutivism may have if it is a good explanation leads us to the second major motivation for developing the constitutivist project. The core idea here is, put simply, that constitutivism seems like it might provide a *good* explanation of the normativity of moral norms.

In particular, if theories about some aspects of agency with constitutivist features turn out to have normative upshots and such theories happen to be true (or at least are very likely to be true), they would *ipso facto* provide an extremely strong vindication of those normative upshots. It is therefore worth exploring action and agency for themselves to see whether good theories about them generate constitutivist upshots.

A third possible set of reasons for going constitutivist comes from first-order ethics. Maybe constitutivism is just morally good or right. For example, perhaps constitutivism is good at quelling first-order ethical worries – e.g. assuming that widespread normative disagreement is problematic, and we want public norms that everyone can hold on to, then constitutivism might explains how the same norms hold for all (Katsafanas, 2018). Or, alternatively, if practical reasons crucially depend on people's subjective and individual motivational sets, then how can we blame people if they cannot be motivated to do what we think is wrong (Williams, 1995)? If some desires are constitutive of all agents' motivational sets, then constitutivism can explain how all agents have the same reasons, and hence can be involved in the same practices.

Similarly, there might be positive first-order moral reasons that indicate that agency works in constitutivist ways. Perhaps we have reason to value our agency as part of a general, independently established, Kantian moral framework. Then that framework could receive a deeper, yet still possibly first-order, justification if the value of is grounded in agency itself.

However, I remain sceptical about these first-order ethical motivations for constitutivism. This is because I share the metaphysical and epistemological worries behind the anti-sceptical motivations for constitutivism. With those worries in mind, I do not trust first-order moral intuitions prior to having a deeper explanation of them – for example, a constitutivist explanation. So, for now, I will work from the metaethical and action-theoretical motivations for the view rather than the first-order normative motivations. The reader may have other theoretical preferences, perhaps by being more trusting about our first-order moral intuitions, but I shall leave it a task for her to go about interpreting my arguments in other ways.

(4) How Can You Really Be a Constitutivist?

Despite the positive motivations for going constitutivist, constitutivism has also started to receive a significant number of criticisms. To defend the view, I will have to reply to them. In particular, four of these objections will structure this dissertation. In chapter 8, I shall summarize how they help me: competitors suffer from all but one of them in interesting ways, whereas the positive view I shall present does not. There are several other general problems for constitutivists, however, that I also will deal with in chapter 8. But I will present more general solutions to these problems, so other constitutivists should be able to accept what I say there.[10]

Of the four main problems, the first is a dilemma that I call the *problem of substantive assumptions*. The more properties constitutivists think that agency has, the more likely it is that they can explain normative phenomena, but simultaneously, their views become less plausible, as more can be denied. On the other hand, weaker assumptions about agency are harder to deny, but it is less plausible that they can be used to explain sophisticated normative phenomena.[11] I spend chapter 2 arguing that all leading constitutivist explanations of morality fail on the first horn. Moreover, fail on the first horn of the dilemma because they suffer from another major problem, namely the *problem of adequacy*. I shall argue that they make inadequate – or false – assumptions.

I shall also make use of the *agency-shmagency problem* (or the *shmagency objection*). The fundamental worry here is that constitutivism seems inadequate at explaining norms because one might not instantiate the aspect of agency that explains norms (i.e. one can be a 'shmagent'). I dedicate chapter 3 to it, though my own solution does not appear until chapter 7.

The final problem I shall use to develop my view is the *problem of bad action*, or the problem of explaining how there can be normatively bad instances of action (or some other aspects of agency) if action is (or those other aspects are) normatively constituted. I will use this problem to defend a particular disjunctivist account of action, where successful actions are a fundamental kind that unsuccessful actions do not belong to. This move will also give my view the right shape to avoid the shmagency objection when I return to it in chapter 7, section 4, and possibly even to respond to the problem of deviant causation for causal theories of action (appendix A, section 2).

[10] Note, however, that my list does not exhaust all issues that have been raised for all constitutivism. For example, Katsafanas (2018) lists some that apply to his interpretation of constitutivism, but not all interpretations. Whether practices or aims can generate reasons depends on whether one takes that to be an interesting question. I do not.

For another example, Bukoski (2016) argues against Smith's framework by challenging its normative upshots. But, as I argued in the last section, I think we should be sceptical about our first-order moral intuitions prior to finding some deeper, e.g. constitutivist, justification of them. Hence, relying on constitutivist-independent first-order moral intuitions to evaluate constitutivism is unhelpful, in my view.

[11] Some constitutivists might even be charged with unstable solutions to this problem, equivocating between the different horns (cf. Katsafanas, 2018; Tiffany, 2011).

However, other constitutivists may also make use of my solution to the problem of bad action. It should be possible to incorporate it in their views. It is, therefore, the first problem for constitutivism to which I will present a general solution. There are also other problems with similarly general solutions. First, there is the *problem of alienation*, according to which it is unclear why we ought to care about constitutively defended norms, even if we are constituted in such a way that we have to care about them. They may be some sort of delusion that we are under. Second, there is the *problem of contingency*, according to which the constitutive norms seem to cover too few possible worlds, and hence do not seem to be necessary, or close enough to that, to be normative. Third, more grandly, there is the *is-ought problem*, or the problem of explaining what is happening when constitutivism seems to wring normative conclusions out of descriptive premises. And finally, fourth, there is the *metaethical problem*, i.e. the problem of locating the constitutivist strategy in the metaethical literature. All these problems will be given general solutions in chapter 8. It follows, I hope, that I will have been able to defend a form of constitutivism by appealing to both its positive motivations and its problem-solving abilities.

(5) Conclusion

To recapitulate: constitutivism says that normative phenomena have force and hold for agents in virtue of properties of features that are constitutive of various aspects of agency. In the rest of this dissertation, I shall focus on moral constitutivism, i.e. constitutivism about normative phenomena that are universally prescriptive and conventionally recognizable as moral.

I shall endeavour to develop such a view in the light of the virtues of constitutivism I have emphasized: it can deal with metaphysical and epistemic queerness, and it stems from independently plausible premises about some aspects of agency. And I shall make use of the problems of substantive assumptions, adequacy, shmagency, and bad action to formulate my preferred version of constitutivism in the rest of the dissertation. Finally, the last few problems for constitutivism will be given general solutions in chapter 8.

2. Moral Constitutivisms

The first of the main problems for constitutivism mentioned in the last chapter is that the more one thinks is constitutive of agency, the less plausible one's account of it risks becoming, for the more there is to be denied. On the other hand, the less one thinks is constitutive of it, the less plausible it is that one can explain something complex – like morality – with it. This is the problem of substantive assumptions.

How do constitutivists grapple with the first horn? In this chapter, I will discuss the most significant contemporary constitutivist theories about moral norms. I will argue that, though they do so in different ways, all leading moral constitutivists fail to make good on the background action-theoretical assumptions that are supposed to be able to explain moral norms. This general failure stems from the inadequacy of their action-theoretical assumptions. So the leading constitutivists in the literature suffer from the problem of adequacy along with the problem of substantive assumptions.

Why this fairly well-known strategy (cf. Setiya, 2007, ch. 2; Williams, 1985, ch. 2, for older versions)? I clearly think there is something promising about constitutivist explanations of morality. Hence, even if the leading constitutivist accounts fail with respect to their assumptions about agency, it might still be the case that some other version of constitutivism can be based on more solid assumptions. So instead of taking the problems for previous accounts to be knock-down arguments against constitutivism, I shall treat solving them as indirect arguments for the view I shall develop myself.

Here, however, I start in section 1 by drawing some distinctions to organize the various constitutivist views. Then, in section 2, I argue that Korsgaard fails to show that actions must be attributed to a fully unified agent, in section 3 that Velleman fails to explain the relation between self-knowledge and control, in section 4 that Katsafanas is wrong to assume that willing power is a constitutive aim of action, and in section 5 that Smith's idealization assumptions do not support ascribing the desires he wants to ascribe to ideal agents. I conclude in section 6.

(1) Preliminary Notes

This is how I defined practical moral constitutivism:

> (*FORMAL CONSTITUTIVISM*$_{PM}$) An account T of some normative phenomenon P_{PM} is constitutivist iff T entails that P_{PM} is normative because P_{PM} is, or is normative in virtue of, some property or properties of the constitutive feature or features C of an aspect of agency $A*$, where C constitutes something as an $A*$.

I shall focus on discussing four leading views that fit this schema. This strategy has two drawbacks. First, there are many constitutivist views that fit the schema, but I do not discuss in depth. Second, there are constitutivists who do not explain moral normativity in the way that I want to do when applying the schema. How shall I treat them?

Regarding the first question, there are at least three forms of constitutivism that I do not discuss in depth. The first type comprises social forms of constitutivism. Some social constitutivists treat some aspects of agency – e.g. rationality, personhood, or agency itself – as in part socially constituted (cf. Berdini, forthcoming; Kosch, 2018; Pauer-Studer, 2018; Walden, 2012; 2018). Alternatively, other social constitutivists run premises that do not generate morality on their own together with premises about the social nature of humanity to generate morality. The latter occurs, for example, on Hobbesian views where maximizing rationality might be constitutive of agency, but that leads to a war of all against all in the state of nature, and a Leviathan is needed to solve it. I suspect that most of these views suffer the problem of substantive assumptions, but I defer discussion of them until chapter 7.

The other two forms of constitutivism I will not treat in depth are Kantian and Aristotelian. While some views of these kinds are very social, there are also many Kantian views that arguably *do* fit my schema even without being social – that is how I read the Groundwork myself (Kant, 1996; cf. e.g. Schapiro, 2001). The same is true for Aristotelian or Aristotelian-inspired Thomist views (e.g. Foot, 2001; Frey, 2019; Thomson, 2008; cf. Korsgaard, 2008; 2009; 2019; Silverstein, 2016 for discussion).

What will I say about these? I think the most significant Kantian and Aristotelian moves also are made by writers I discuss. As they are the most well-developed forms of Kantian constitutivism in the literature, I shall treat Korsgaard's and Velleman's views as representative of Kantianism (in sections 2 and 3 respectively). Similarly, the Aristotelian views are not very strictly developed *qua* versions of constitutivism. Instead, insofar as Aristotelian premises are used in constitutivist arguments, they tend to be put to non-Aristotelian normative purposes (by e.g. Korsgaard and Smith). So insofar as I discuss these views, my criticisms will challenge various ways of developing Aristotelian constitutivism.

The second main drawback here is that I will not discuss all constitutivist views, but only those that explain the normativity of morality *in the right way*. There are other constitutivists who try to explain moral normativity, but not in line with my universal prescriptivity *desideratum*. Most notably, such authors include many Humeans, such as Sharon Street (2008; 2012) and Jamie Dreier (1997). But many valiant attempts to explain morality without universally prescriptive notwithstanding, I still do not think such views manage to explain morality – they explain universal prescriptivity away rather than positively, so it ultimately ends up boiling down to contingent desires or commitments that some agents may lack. However, since I take it to be true that morality indeed is universally prescriptive, I shall aim to discuss theories according to which universally prescriptive moral normativity can be positively explained.[1]

(2) Korsgaard

Christine Korsgaard has developed what probably is the most influential version of constitutivism so far. To that end, she has presented several arguments, usually of fairly orthodox Kantian varieties. I shall, however, mostly focus on her argument in Korsgaard (2009), where she suggests that agents need to unify themselves to act, and that this implies that *CI* is constitutive of agency.

To show why, I will need to take a detour by way of some of her other arguments. Earlier, inspired by Kant's argument in Groundwork III (Kant, 1996, pp. 94-108), in Korsgaard (1996*a*, pp. 97-98; cf. Korsgaard, 2014), she re-formulates his argument for thinking that rational and free action are the same to claim that that the source of normativity is the autonomy of the will. Put extremely briefly, she claims that a free will is a 'rational causality which is effective without being determined by any alien cause' (Korsgaard, 1996*a*, p. 97), viz. by anything apart from itself. But since it is a cause (i.e. 'a kind of causality'), it is law-bound – though not by anything else than itself. Hence, it could only be a law that commands that one acts from the form of law itself, i.e. *act only*

[1] For the same reason, I will not spend much time on the non-morality-explaining norms that some of the more prominent morality-explaining constitutivists also discuss or endorse, such as norms of coherence (cf. Korsgaard, 1997; 2009, pp. 68-72; Velleman, 2009, ch. 4-5; Katsafanas, 2013, pp. 48-53; Smith, 1994, ch. 5; 2009).

in a way such that your maxims could be made into universal law; or, in other words, the formula of universal law (*FUL*) version of *CI*.

She also adds that the faculty of the will is practical reason, and action by will is necessarily based on reasons. These reasons are, however, encapsulated in principles, so the will must have a principle. But, just as a free will cannot be determined by anything else, the will, too, must be self-governed – or, using another word, autonomous. So the principle of a free will cannot be based on anything outside itself. Again, it follows that one can act freely only on a principle which has the form of a law and nothing more, and hence governs itself. Hence, free action and action for reasons are united under *FUL*. As *FUL* only requires us to act under the form of law-boundedness, *FUL* is the principle of free will or practical reason.

Versions of this argument have been controversial since Kant's original formulation. One may, for example, have worries about the Kantian conception of freedom – though see Korsgaard (1996*b*) – or about the form of law that Kant and Korsgaard think an action must have, or about whether reasons must be encapsulated in principles. Why any of these things would work in Kant-friendly ways is unclear, and any significant progress here seems unlikely to come swiftly. Hence, because these assumptions are so substantive, this argument suffers from a version of the problem of substantive assumptions. Weaker premises would, *ceteris paribus*, be more attractive.

Korsgaard also gives us an argument for the formula of humanity (*FOH*) interpretation of *CI*, which is supposed to provide a clearer delineation of which maxims we are to universalize by *FUL* (Korsgaard, 1996*a*, pp. 122-125; cf. Kant, 1996, pp. 79-80). Here, Korsgaard claims that actions require that we put our practical identities behind them, but we can step above any of our particular identities and question whether it gives us reasons. Taking up that reflective stance is to be human, so action requires identifying as human. And as valuing at all implies valuing oneself as valuing the action, we must value our own humanity. And this valuation generalizes to others, not least because there is no difference between our reasons and others'; reasons are public, not private (Korsgaard, 1996*a*, pp. 132-145).

However, I share Street's (2012) suspicion that the premise that valuing implies valuing oneself is implausible. A community of creatures – Street's analogy is bees – who only value others in the community, not themselves, seems entirely conceivable. Again, it would be helpful for Korsgaard to work from less orthodox Kantian premises; this argument wears its problem of substantive assumptions on its sleeve, and is most likely inadequate in the light of the bee community counterexample.

We shall have to look elsewhere for an argument with better premises. This takes us to the argument that I mentioned that I want to focus on. The argument in question is a new argument for thinking that the *FUL* version of *CI* is constitutive of action (Korsgaard, 1996*a*, pp. 225-233; 1999; 2009, pp. 72-80). Korsgaard calls this argument 'the argument against particularistic willing', and I shall primarily be concerned with the formulation in Korsgaard (2009). There, she uses it – together with some extra premises about human sociality – to ground the other versions of *CI* as well, so the argument is a central premise in her more systematic view.[2]

The argument against particularistic willing can be broken down roughly like this. First, to act, an agent performs a procedure by which she makes herself into a cause in the world. Second, to make herself a cause in the world, the agent makes herself into a unified whole. Third, she unifies herself if (and, presumably, only if) she acts from universalized Kantian maxims. Fourth, she has universalized maxims if (and, presumably, only if) her actions are approved by *CI*. As universalized maxims are constitutive of agency, *CI* is constitutive of agency.

Some clarifications are needed before I present the argument in more depth. First, I will be talking about human rather than animal actions or agency when discussing Korsgaard's framework. Animal actions work differently on her picture – they are much more instinct-based (Korsgaard, 2009, ch. 5; 2018, ch. 3-7).

Second, a 'maxim', for Korsgaard, is an explication of an act-type in which a means is taken to a specified end. A typical example would be 'To grab a coffee, I go to the kitchen.' Grabbing a coffee is my end, going to the kitchen is the act. Maxims are, Korsgaard thinks, identical to Aristotle's *logoi* – i.e. principles – so maxims and principles are, fundamentally, the same kind of things (Korsgaard, 2005).[3] Principles describe the agent's taking certain considerations to count in favour of actions, so the agent chooses maxims using her principles.

[2] More specifically, she appeals to her theory of public reasons and the nature of social interaction to argue that we must respect the humanity in ourselves like that of others, that we constitute ourselves over time by respecting ourselves, and that the respect places us in a kingdom of ends (Korsgaard, 2009, pp. 204-206). These extensions rely on having established *FUL* from the start.

[3] Moreover, she adds that our principles, and especially our constitutive principle, are akin to our Aristotelian forms. Hence, we are the kinds of creatures we are in virtue of having our forms (cf. Korsgaard, 2009, ch. 1; 5-6).

Third, Korsgaard takes *CI* to be the deepest principle of human action. *CI*, in the form of *FUL*, is used to evaluate all other maxims because they all have to be universalized by it. It also has a universal form that the others lack because it covers all other actions, whereas contingent ends are more limited. In fact, while human action also is governed by the hypothetical imperative (*HI*), according to which one is rationally required to take means to one's ends, *CI* fundamentally incorporates even *HI* (Korsgaard, 2009, pp. 71-72; cf. Korsgaard, 1997).

Korsgaard's argument, then, goes like this:

(1) If an agent is to perform a human action, then she precedes her action with a procedure by which she makes herself into a cause in the world.
(2) If an agent precedes her action with a procedure by which she makes herself into a cause in the world, then the agent is a unified whole.
(3) If the agent is a unified whole, then she wills according to universalized maxims.
(4) If the agent wills according to universalized maxims, then her constitutive principle is *CI*.

(C) If an agent is to perform a human action, then her constitutive principle is *CI*.

How should we unpack these premises? Korsgaard does not discuss (1) at much length in her (2009).[4] However, in later work (e.g. Korsgaard, 2014; 2018; 2019), she has emphasized the argument here much more explicitly, and some work is needed to explicate it accurately. Fundamentally, Korsgaard thinks, just like Kantians hold that one must be able to infer that there is a thinker behind every thought, there must also be a doer behind every action (Korsgaard, 2018, p. 34). And we can infer that because the agent precedes her action with a procedure by which she makes herself into a cause in the world.

In her 2009, Korsgaard's main consideration in favour of this conclusion appears when she argues that *HI* is a constitutive principle of willing (Korsgaard, 2009, pp. 68-70). Her point is that to act is to initiate a causal chain running from oneself and into the future, and to be at least somewhat successful doing so. But one cannot do so unless one does something to insert oneself as a cause into the world, so *making* oneself into a cause

[4] The 1996*a* version, however, is premised on Kantian claims about causation contrasted with succession.

is what differentiates willing from merely desiring, wishing, or similar. Hence, she thinks, one must take a means to an end to will anything (rationally).

This argument stems from the same intuition as the more common argument from the disappearing agent against event-causal theories of action. There would be no will – and no agent – behind the action, Korsgaard thinks, unless the agent makes herself into a cause behind it, for then the agent would disappear.

But what does it mean to say that one makes oneself into a cause behind an action? The problem of the disappearing agent has been formulated in many ways by many authors, but Korsgaard thinks there are two problems here (Korsgaard, 2014).[5] They are both, in some sense, deeper than the questions about whether actions are free, since they apply to both free and unfree actions. The first one is causal and descriptive; it is unclear how an agent actively initiates an action rather than becomes reduced to yet another event in a causal chain when her actions are causally efficacious – and actions are indeed, Korsgaard thinks, causally efficacious. More metaphorically, we should be able to show how agents *author* their actions. They do not necessarily author their actions if they themselves are just caused by (and, in some sense, 'authored by') other background causal chains.

Though Korsgaard is not wholly explicit, I believe we can explicate the problem like this. Whatever one thinks about free will, there are two minimal features that any theory of action or agency must capture to explain how the agent does not disappear. The first is that the agent must *stand behind* her action, so that she is something over and above it. It is from this position that the agent may issue in her action causally efficaciously. Moreover, the agent must *actively* bring the action about, in a sense of 'activity' where there is something that the agent does to initiate an action rather than letting it happen to, or through, her, but where that activity need not be an action itself. Regardless of what else the causal and descriptive aspect of the problem of the disappearing agent is, it involves *at least* these two features.

[5] Beyond Korsgaard, some authors have raised it in the context of discussing free will (Melden, 1961, pp. 128-129; Nagel, 1986, pp. 113-114). Melden's worry is that the agent seems to become too minimal to stand behind an action, but other authors worry that the agent disappears completely on an event-causal picture (Hornsby, 2004). For present purposes, I am concerned with the latter issue of how the agent risks disappearing completely – I do not really see why it is a problem if the role of the agent is minimal (cf. chapter 5, section 3; 4).

Notably, also, for Velleman (1992), the problem is twofold, and concerns both how an action can be originally authored and sustained by the agent over time. I believe that an answer to the authorship question pretty straightforwardly carries over to the question of sustaining it, so I shall ignore this extra complication.

Furthermore, Korsgaard also thinks that the problem of the disappearing agent has normative import. More specifically, she thinks that agents can succeed or fail in ways that ordinary causes cannot, and that we must be able to respond appropriately or inappropriately to actions because agents are responsible for them, and hence agents must be involved in their actions.

Her core solution to the problem of the disappearing agent, featuring all these descriptive and normative aspects, is to say that the agent has to precede her action by performing a certain psychological procedure. This procedure involves a kind of activity by which she authors her actions. In Korsgaard's case, as we shall see, the activity is to unify herself behind a universalized maxim, and hence to be active in virtue of acting on a principle – *CI* – by which the she universalizes her maxim. Because the agent contributes to her action with a kind of unification that stems from her universalizing her maxim, there is something that the agent does to author her actions. Hence, the agent stands behind her action by doing something active (and she can succeed or fail doing so).[6]

On to premise (2), according to which an agent has to unify herself to perform the right kind of procedure to stand behind her action. The deepest reason for accepting the premise is that, in action, we identify with our principle of choice, so we take ourselves to be that principle. Therefore, Korsgaard thinks, we have the phenomenological experience of being torn apart when we have contradictory impulses that may issue in different maxims (or principles).[7]

[6] Korsgaard also uses the normative aspect of the problem of the disappearing agent to argue that agency is identical to personhood. Her Platonic psychology (Korsgaard, 2009, ch. 6-7) appears as a deeper explanation of this point. I have relegated it to a footnote because it will play no further role in my discussion of her view.

Nevertheless, the view is worth recapitulating. Korsgaard thinks that agents' identities are constituted by the maxims they choose, but deeper down, they all have to be regulated by *CI*. Here, Korsgaard defends what she calls a constitutional model of the soul (or 'person'). She thinks we can explain action in general – and how it is attributable to its author as a whole – by a Platonic conception of how a constitution is something over and above the citizens of a state acting individually. There are three aspects that work together here: appetite, proposing desires to the agent, reason, deciding whether to act based on these, and spirit, following through on the decision. They work in the same way in a person as in a city-state, Plato famously thinks.

The Platonic picture maps onto Korsgaard's Kantian one. We experience desires, we deliberate about whether or not to act based on them (when choosing according to principles), and then carry out our decisions. And we, as agents, do not cause an action until the entire procedure, settled by our constitution, is completed. For Korsgaard, then, we are identified with our constitution, which is *CI* (in the form of *FUL*).

[7] Korsgaard also defends premise (2) in some other ways. First, she seems to deem the entire inference from (1) to (2) a conceptual truth – to attribute an action to an agent, the agent must be a unified whole. Second, she makes the phenomenological point that, as we identify with our different principles of action, insofar as we do not know what to do, we feel torn between different identities (Korsgaard, 2009, pp. 125-132). This point could support (2) independently of the deeper explanation she gives it.

We do that, she adds, because we act for maxims (which are identical with principles). With that point in mind, she argues by a dilemma. Either we identify ourselves with our principle of choice, or we do not. If we do, we must unify ourselves when acting, but if we do not, the principle becomes a third thing within us, comparable to the desires that pull us towards different possible courses of action. If so, we become passive observers of our mental states rather than agents.

But when we act, we are not passive like that. We regard the principles by which we choose to take some course of action as expressive of ourselves. We necessarily do so, for if the principle by which we choose between the courses of actions supported by different desires were not, then the principle would be another causal force working inside us. In other words, we would be part of the event-causal chain that we must stand above to act. But then we cannot regard the movement that it might issue in when we choose to act as our action – it is a mere event, happening through us (Korsgaard, 2009, p. 75). And hence, we must identify with the principles of our choice, so we unify ourselves by acting from our principles.

Premise (3) explicates what unification involves. In Korsgaard's sense of the word, it proceeds by universalizing one's maxims for action. An agent wills *particularistically* iff she thinks she can use a maxim once (or an arbitrary set of times) and then discard it, without needing to change her mind about whether or not she may do so based on some good reason. Universal maxims, then, are any maxims that are not particularistic, either in the sense that they are absolutely universal, so we never will change them, or provisionally universal, so we never will change them *ceteris paribus* (Korsgaard, 2009, pp. 73-74). This means that we always will accept them, in the sense that they apply in all possible situations, future or present, unless there is good reason to change them (Korsgaard, 2009, pp. 78-79).

Particularistic maxims cannot, however, be ours. As Korsgaard argued when defending premise (2), there is something that is us that stands behind every action, and we identify ourselves with. This something is our unifying principle. The second aspect of that point, yielding premise (3), is that if we were to will particularistically, we would eradicate the distinction between us and the action, or, in other words, between our unifying principle and the action. We would no longer be something over and above the particular action, so there would no longer be an agent behind the action, causing it. This is because the universalized maxim always covers more possible situations than the present one, while particularistic ones always can be discarded. But if they can be, then so can our agency, so our agency becomes a causal force that need not stand behind and issue in our action(s). Hence, we cannot will particularistically, and therefore we must will through universalized maxims.

Premise (4) is the concluding premise in the argument. Korsgaard seems to think that a maxim is universal (only) if *CI* is our constitutive principle, i.e. the principle that most deeply makes us into who we are. For giving *CI* that role is the same thing as having a universalized maxim; *CI*, in the form of *FUL*, requires us to act only on the maxims that we could make into universal law. Hence, an agent is unified when constituted by *CI*. The conclusion, (C), follows.

The argument against particularistic willing supports a form of moral constitutivism. It fits my schema from chapter 1, section 2, unproblematically. The reader may remember that it has several features: *T* (the account), P_{PM} (some moral norm), *C* (properties of the constitutive features which explain why a norm is normative) and *A** (some aspect of agency). Clearly, Korsgaard's view *T* involves an attempt to explain the normativity of *CI* (P_{PM}), as part of an explanation of practical normativity in general. The core idea, then, is that *CI* (P_{PM}) is constitutive (*C*) of human agents (*A**). And as agency is 'plight inescapable', meaning that we are continuously faced with new situations where we must act by *CI*, *CI* has a normative grip over us (Korsgaard, 2009, pp. 1-2; cf. chapter 3, section 2; chapter 5, section 6, below).

Moreover, *CI* in the form of *FUL* is not an aim of action, but a principle in Katsafanas' sense (cf. chapter 1, section 1). But it still seems to be a moral norm. It is universally prescriptive because it is constitutive of all human agency. And it is easily recognizable as having moral import, so it is recognizably moral.

There is a potential complication here, however: Korsgaard thinks that the *FUL* interpretation of *CI* only is one part of the story about morality. In its fundamental form, Korsgaard thinks, *FUL* is a principle of practical reason, and *FOH* is needed to give it content.[8] To that end, as I mentioned in footnote 2, she also argues for the respect for humanity and kingdom of ends interpretations of *CI*, providing more substance to her

[8] Clarifying her views, Korsgaard says: '[f]irst, there is [*FUL*] in the purely formal sense, [which in (Korsgaard, 1996*a*)] I called [*CI*]. Then, there is [*FUL*] in the sense of a principle, which demands that we act on reasons that we can share with all rational beings who live together in a cooperative community, which in [Korsgaard, 1996*a*] I called the Moral Law. And then, there is the [*FOH*]. In [Korsgaard, 1996*a*] I argued that the foundational argument in fact only gets us to the formal version of [*FUL*], [*CI*]. Before we can know what that principle requires of us, we need to specify what the principle universalizes over. The argument [for *FOH*] shows us that what [*FUL*] universalizes over is human beings as ends in themselves. So by that route what the foundational argument actually brings us to most immediately is something more like [*FOH*].' (Korsgaard and Pauer-Studer, 2002, p. 11)

take on morality. Nevertheless, *CI* even in its *FUL* form lays the groundwork for the rest of her arguments, and it is also obviously conventionally recognizable as moral by itself. So I shall focus on it.

Being arguably the most elaborate constitutivist in the literature, many of Korsgaard's claims have already been subject to much discussion. Here, I shall focus on premise (3).[9] Elijah Millgram has denied it, saying that Korsgaard has presented a false dichotomy between willing particularistically and willing universally. Perhaps an agent needs to unify herself to some degree to act, but hardly in a sense as strong as Korsgaard's (Millgram, 2011). It seems enough to have a policy, a commitment, a long-lasting desire, or something else along those lines to stand behind any particular action; they all cover more cases than a single particular one, so there is no need to have universalized maxims over all similar present and future situations. I endorse this worry.

There are other worries about premise (3) too. While I did claim that the argument against particularistic willing seems better than Korsgaard's main (1996*a*) arguments, there are lingering questions about the Kantian premises in this argument too. One problem is that Korsgaard's universalized maxims seem very theoretically inelegant when it comes to explaining action. Ordinary action ascriptions do not need to involve universalized maxims – as a matter of how we talk about what we do and why we do it, we do not need to explicitly cite maxims, just the reason they supposedly encapsulate. So, *ceteris paribus*, it might be better to do without them.

There are two sub-issues here. First, maxims are supposed to embody act types, but – theoretical assumptions aside – we have no obvious folk-psychological reasons to cite act types rather than tokens in most action explanations. Folk-psychologically speaking, it is enough to have one particular desire or reason to do something to act based on it. Because the maxim machinery becomes an additional theoretical add-on, it might be better to do without it in action explanations.

Second, why would we need Korsgaard's type of universality? She grants that maxims may be highly specific while still being universalized (2009, p. 73). Fair enough. But nothing about ordinary action explanations suggests that agents need to act on the same maxim in all similar present or future situations, even *ceteris paribus*. More would

[9] I say this while recognizing that there may well be problems with the other premises too. For example, one problem for (4) is that it is unclear whether Korsgaard can make good on how we act for *CI* rather than just in line with it. It is one thing to say that we need to have universalized maxims, but it is quite another to say how they are formed. Perhaps we are just arbitrarily interpretable as having maxims that are in line with *CI*. Korsgaard needs to say more here.

have to be said to defend that assumption in explanations of particular actions.[10] In both these cases, then, Korsgaard's machinery seems extravagant. It would be preferable to do without her assumptions, if possible.

One response on her part would be to suggest that maxims are necessary to rationalize actions. If we are such that we randomly move about on our desires without universalizing them (in her terminology: are 'mere heaps'), it is not clear how our actions are intelligible to us or others. We seem to be moved by our desires, not our wills. And as action explanations are supposed to rationalize our actions, universalization may still be necessary.

However, I fail to see what it is about φ-ing on a whim that makes the φ-ing unintelligible or non-rationalized. Expressions like 'on a whim' or 'because I wanted to' or 'because I felt like it' suggest that we do not need universalization for rationalization. We often explain action to each other just by these minimal notions (cf. appendix A, section 1, for more discussion).

A more general response to all my critical points would be to stick to Korsgaard's defence of premise (3) and go *modus ponens* from that, even though I am inclined to go *modus tollens* based on the challenges above. Essentially, that would involve denying my responses, because Korsgaard can argue that the underlying thought that there *has* to be an agent behind the action, who is not instantiated in just a particularistic maxim, is right.

But denying the worries about her Kantian assumptions seems inadequate. This is because denying that universalized maxims must play a role in actions goes in tandem with Millgram's suggestion that there are other very plausible explanations for why she does not need to universalize. As Millgram suggests, the agent can stand above any particular action in virtue of holding on to other mental states, such as policies, commitments or long-lasting desires. This means that (3) just seems false, hence making a *ponens* move implausible.

Another reply relies on reformulating premise (3). Perhaps one could go with:

> (3') If an agent is to become a unified whole, she is regulated by
> an ideal of universalizing her maxims.

[10] I suspect that Korsgaard, like Kant, wants universalization here because she follows Kant in thinking that action explanation takes the form of causal laws (cf. Korsgaard, 1996*a*, pp. 97-98). But just how they work, or why they are relevant here, are new cans of worms, so the problem for her view deepens with this assumption.

If (3') is accepted instead of (3), then maybe one can handle the worries above. An *ideal* agent would maybe have universalized maxims (as per *CI*), because one may interpret universalization as a normative claim. On this view, the idea is that the agent *ought* to universalize her maxims to act, rather than having to be such that she actually universalizes them to perform every action.

But this move is not open to Korsgaard. We learn this from cases of failed action. An agent who does not act on a universalized maxim because she fails to be regulated by an ideal of becoming a unified whole is steered by heteronomous forces in a way that is exactly analogous to what Korsgaard attempts to rule out when defending premise (3) – if the agent is not actually unified, she does not stand over and above her particular action, so her will is dominated by some external force. Hence, if becoming a unified whole is a normative ideal that we fail to live up to, there is no 'us' behind our actions, meaning that we cannot infer that the agent wills (or aims to will) by universalized maxims. I conclude that even the argument against particularistic willing fails.

(3) Velleman

David Velleman's form of constitutivism is based on the idea that self-knowledge is the aim of action. Like Korsgaard, he has defended his view in many guises, and as was the case with her view, I shall discuss some alternative interpretations of the view before I arrive at the most plausible one.

Sometimes, Velleman (e.g. 1996; 2007*b*) has suggested that an agent is constituted in part by a desire for self-knowledge or self-understanding. On another version of his view, *drives* to self-knowledge or self-understanding generate these desires, where drives consist of some form of general psychic energy, comparable to a Freudian libido (cf. Velleman, 2007*b*, introduction). And sometimes his formulations suggest that both desires and drives may be operative (e.g. Velleman, 2007*a*; 2009, pp. 15-19; 106-109).

However, these versions of his view seem *prime facie* implausible. The motivational states that aim at self-knowledge seem problematic for the same reason that Korsgaard's Kantian assumptions do. In many folk-psychological cases, we do not refer to them – for example, it seems contrived to attribute desires or a drive for self-knowledge in cases where an agent does something simple such as going to the kitchen to make a sandwich because she feels hungry (Hazlett, 2009). So there is a presumption against formulating Velleman's view so that it is based on desires or drives; it seems to suffer from the problem of substantive assumptions, and is probably also inadequate if one formulates it in that way (because it is based on unwarrantedly strong assumptions).

Admittedly, Velleman's main case for desires for, or a drive aiming at, self-knowledge can be construed as abductive. He does not just try to capture folk discourse; he attempts to solve general problems in action theory as well as providing independent – even empirical – support for the view (Velleman, 2000). This is a laudable strategy, and it makes his view very hard to conclusively refute.

But it is, simultaneously, very hard to conclusively show that it is the best explanation of action or agency. The Humean conception of agency I shall defend does not include a motive of self-knowledge and simultaneously solves some of Velleman's problems, such as the shmagency objection I shall charge him with in chapter 3, the problem of the disappearing agent (cf. Velleman, 1992), and the problem of providing an explanation of morality (cf. Velleman, 2009; 2013). It is unclear whether Velleman's view would be better than mine in general.

Even so, another interpretation of his view seems more palatable. We should, presumably, attribute *intentions* to agents even when they perform simple actions. And in what strikes me as the best version of Velleman's view – which Velleman at least sometimes seems to have thought himself too (2000, pp. 15-24; 2007*a*) – the self-knowledge that agency aims at is explained as a feature of intentions. Intentions are, in turn, understood as a kind of self-directing mental states that count as a kind of beliefs, which, when successfully enacted, amount to self-knowledge. I shall discuss this version of his view, and how it might explain morality, in more depth.

Velleman's aim here is to explain how full-blooded action differs from non-action activities. With the Anscombean idea that acting agents have a special kind of self-knowledge in mind, Velleman expands on the theme of holding truth to be a constitutive aim of belief (Shah and Velleman, 2005). The key point of his explanation is that beliefs can be (or can *become*) true in different ways. Beliefs are ordinarily made true by corresponding to the facts. But some beliefs make the facts correspond to them instead.

More specifically, Velleman's claim is that beliefs about oneself, in action, make themselves true by causing the world to conform to them.[11] The argument proceeds from

[11] At times, he suggests using other terms than 'beliefs' to demarcate these states. For example, he associates them with 'expectations' (Velleman (2000), p. 24), 'choices', 'fore-thoughts' and 'self-understandings' (Velleman, 2009, ch. 1; 5), and in his (2007*b*, xix), he is explicit about not wanting to use the term 'belief', preferring instead to say that an intention 'represents its content as true with the aim of doing [what the content would have one do] if the content really is true.' But I shall stick to 'belief' for simplicity. Nothing hangs on the terminology.

the claim that actions have two different kinds of aims. There are first-order aims, which are one's particular goals. But there are also higher-order aims. Such aims are activities one engages in to reach one's first-order aims.

Particular aims all partake in the same activity of being consciously pursued by the agent. The activity of being consciously pursued, one's 'having a *controlling consciousness* of one's behavior, a guiding awareness' (Velleman, 1996, p. 720, his italics), is a person's way of exercising control over her actions. Velleman identifies having such a controlling consciousness of an action with having a special kind of self-knowledge of what one is doing in that action (Velleman, 2000, p. 22). It follows that when one acts to (try to) achieve a particular aim, one has the higher-order aim of self-knowledge. This higher-order aim, moreover, should be identified with an intention. And an intention, in the primary sense of the word, is a particular kind of belief (Velleman, 1996, footnote 55; 2000, pp. 21-22). When true (and reliable), it yields a kind of self-knowledge which is 'directive rather than receptive' (Velleman, 1996, p. 720).

It is here that it matters that there are two ways to become true for beliefs: '[o]ne way is to [receptively] accept a proposition in response to its being true; the other is to [directively] accept a proposition in such a way as to make it true' (Velleman, 1996, p. 721).[12] Receptive beliefs (or receptively accepted propositions) are what we usually take beliefs to be: they just aim to represent the world truly. However, when one holds a directive belief (or directively accepts a proposition), one accepts it in a way which pushes one to make it true. If an agent directively accepts that she will φ, this acceptance is what pushes her to φ, hence fulfilling her belief.

Directive beliefs have three key properties that work together to give them their special role. First, they aim at representing some state of affairs as true. They aim to settle questions, not just represent a state of affairs as something to arrange. Second, directive beliefs are truth-functional. Directive beliefs can be countered by evidence that one will not act in the way one originally had envisaged. Third, most importantly, they are desire-like in their direction of guidance. They push one to make their content true instead of aiming to conform to the world.

Because full-blooded actions are (partially) constituted by their intentions, action constitutively aims at self-knowledge. And this self-knowledge, Velleman thinks, explains our moral codes. The codes are shared attempts at regulating this kind of creation of self-

[12] I use the phrases 'directive belief' and 'directively accepted belief' (and vice versa for 'receptive') interchangeably. In Velleman and Shah (2005), Velleman claims that accepting a proposition as true is just what it is to have a belief. On that view, there should be no interesting difference here.

understanding intersubjectively. Our actions must make sense to each of us, in the sense that we have to understand them, or else we cannot interact. We are spontaneous improvisers, settling on various scenarios to enact. But we also have a 'universalizing tendency' (Velleman, 2009, p. 48) towards self-knowledge, insofar as we, as social beings, become more and more integrated into shared patterns of social interaction. What makes sense for us becomes more and more universal, transparent, and mutual in our regard for each other over time, even though these properties need not be universal or universally prescriptive (Velleman, 2009, ch. 6; 2013).

Velleman, then, uses his theory of action to account for morality. It explains the emergence of moral norms. How does this view fit into my constitutivist schema? Velleman's account T involves an attempt to explain our moral practices in terms of shared codes of making sense that arise in virtue of our general aim to know ourselves (P_{PM}), where making sense of ourselves is constitutive of action (A^*).[13] More specifically, the properties of the constitutive feature C of action that do work here are the higher-order aim of self-knowledge, which explains why these norms hold for us, and the fact that the aim of agency is 'naturally inescapable', which it is because it follows from natural human capacities and because it is self-vindicating, as questioning it involves making use of it (Velleman, 2009, pp. 136-142).

Is this even morality, in my sense? Our moral practices are, of course, recognizable as moral. But universal prescriptivity is not needed as a feature of our standard moral practices on Velleman's picture – we can have very different forms of self-understanding, and hence different norms that will be prescriptive relative to them (cf. Velleman, 2013).

The key to understanding how Velleman's view captures morality, in my sense, is to evaluate it holistically. The aim of self-understanding in action is universally prescriptive – everyone has it, and it is naturally inescapable – and in virtue of its universalizing tendency, it explains the emergence of moral codes that are recognizably moral. Universality, transparency, and mutuality are familiar moral themes, and even more so if they are given the Kantian interpretation Velleman prefers (cf. Velleman, 2009, ch. 5; 6). Hence, while the moral norms Velleman explains are not by themselves constitutive of action, their emergence is explained by the constitutive aim of action. And as ordinary

[13] Notably, Velleman simultaneously uses this constitutivist account of action to develop a theory of reasons as rationales (for self-conceptions), where the reasons are whichever conceptualizations there are of bodily movements that are veridical and make sense to the agent enacting them (cf. Velleman, 2000, pp. 27-28; 2009, ch. 5).

moral norms are based on the aim of self-understanding, *that* aim acquires a central role in ethical theory on Velleman's view. The aim has a kind of first-order ethical status, for self-understanding does not just cause moral norms, but also vindicates them as norms generating self-understandings are vindicated by the aim of agency. It follows that folk-recognizable moral codes are based on a universally prescriptive aim which is recognizable as moral from the perspective of first-order ethical theory.[14]

Alas, there are many problems right on the surface of Velleman's complex view. For example, does action require self-knowledge at all (Doris, 2015)? And are these directive beliefs not *too* similar to problematic besires? I shall, however, focus on another, less obvious, problem. This problem is that, whether or not her actions are successful, an agent's self-knowledge does not seem to be closely connected to the control she has over her actions. It may be true that actions both require a kind of agential control and that one necessarily knows what one does when one is acting (though I have my doubts about the second claim). But these features need not hang together, and that indicates that Velleman's view of intentions, which is supposed to capture both, is implausible.

Consider a case. My first-order aim is to enter my house. Perhaps I just desire to go home to have dinner. My higher-order aim, then, is to control my action so I can go have dinner in my house. Does it follow that I *know* what I am doing when I enter my house? Not necessarily. As it happens, the lock on my front door is exceedingly complex to unlock, and I do not know what I am doing when I try to unlock it. I do know, in general, that I tend to be successful doing so. But that is a background piece of knowledge that assures me that I have control of what I do even when I do not have directive knowledge of the control that I have. In Velleman's terminology, it is a receptive, not directive, piece of knowledge.

Or consider another case. Adina is writing her transfer chapter. (A 'transfer', in British academia, is a process a first-year PhD student must pass to start her second year, and a 'transfer chapter' is a dissertation chapter that is supposed to be examined as part of the process.) Her particular aim is to pass, so she can proceed to start her second year as a PhD student. Her higher-order aim is, like mine, to control her action, but in this case to pass the transfer rather than to go have dinner. Does she *ipso facto* directively know that she is writing her transfer chapter, in Velleman's sense of having an intention that

[14] It is also arguable that Velleman's view has an even folksier moral commitment, for 'Know thyself!' has functioned as a normative imperative in various cultural traditions, even though it hardly is a moral imperative in most leading contemporary philosophical theories. I am not sure about whether self-knowledge is conventionally recognizable by the folk as *moral*, however, so I have relegated this point to a footnote.

she is in the process of making true, without necessarily having made true yet? Not really. She could end up submitting another chapter to the examiners, so her intention need not amount to Velleman's kind of self-knowledge – *whether or not* her belief is true.

Perhaps Velleman might reply that I and Adina do not know what we are doing; we only believe that we know what we are doing. But he cannot take this route. We *do* seem to control what we are doing in these cases, but our self-knowledge seems detached from the control that we have – because we do not, in fact, have knowledge of what we are doing in a way that is closely connected to how we control what we do. In my case, I do not know what I am doing when I unlock the door, yet I still seem able to control doing so. In Adina's case, she may even be wrong about what she is doing, because she may end up submitting another chapter for her transfer examination.

Another response is to say that what we fail to know, but still control, is a sub-action which features in a larger action of which we have knowledge. Perhaps I know I am going home, and Adina knows she is working on her transfer material, but we do not know what we are doing in our sub-actions. But we still have some more overarching kind of knowledge.

However, the cases can easily be formulated so that we should know what we are doing more precisely. Perhaps my sole aim is to unlock my front door, I do know that I can open it, but I still do not know how I go about doing that when I do it (in Velleman's sense of directive knowing). Or perhaps Adina's sole aim is to write chapter 4, but, as it turns out, chapter 4 will be called chapter 6 later on in her writing process. But she still seems to control what she is doing when she writes it. So control and self-knowledge come apart. So does even the most plausible version of Velleman's view.

(4) Katsafanas

On Paul Katsafanas' view, action has two constitutive aims: *agential activity* and *power*. From agential activity and power, he tries to develop a Nietzschean normative view according to which agents have reasons stemming from these very aims. Here, I shall criticize the aim of power, which is the aim that does most of the normative work in his framework.[15]

What does Katsafanas' view involve? First, 'agential activity' is a kind of equilibrium; an action is active just in case the agent performing it approves of it now or would keep approving of it when given more knowledge about its etiology (Katsafanas,

[15] In Leffler (2016), I discuss both aims in more depth.

2013, ch. 5). The second aim is 'power', according to which actions aim at overcoming resistance, and are better the more resistance they overcome (Katsafanas, 2013, ch. 6).

According to Katsafanas' interpretation of constitutivism (cf. chapter 1, section 1), constitutive aims ground standards of success, which in turn ground reasons to live up to them. And as one aims at agential activity and power, one has reasons to live up to these aims. In fact, Katsafanas thinks, the reasons provided by one's aims of agential activity and power are such that other values must be assessed in their light.

To this end, the two aims work in tandem. By her agential activity, an agent can approve of her actions to various degrees, depending on what she knows about the actions. And an agent approves, or disapproves, of actions in virtue of her values. By themselves, these values are to be assessed by the aim of willing power. Values ought to be discarded either if they conflict directly with it or might lead to conflicts with it in the future, as the other values then serve to undermine power (in particular cases or in general). Indeed, in the long run, though presumably not necessarily in any given case, the reasons for action that stem from the aim of power win out when compared to other kinds of reasons an agent may have (based on her other values). The reasons that stem from the aim of power are ubiquitous, pervasive, and are (typically) reinforced by other motives, and hence tend to outweigh other reasons over time (Katsafanas, 2013, pp. 39; 191-200).

These two aims of action ground a Nietzschean first-order ethical view, Katsafanas thinks. The constitutive aims of action function as reason-giving ideals that all other values can be weighed against. They are, therefore, able to ground one class of universally prescriptive reasons for action. Even though we to some extent have different values, these are measured against the constitutive aims, as they can realize them (or be inconsistent with them) to different degrees. Katsafanas calls this view *parametrically universalist*, meaning that everyone has the same fundamental ideals – activity and power – though they can be realized in different ways, and to different degrees, for different persons (Katsafanas, 2013, pp. 211-218).

How does the view fit my constitutivist schema? On his account *T*, Katsafanas explains moral norms (P_{PM}) in terms of the aims of activity and power (*C*), which are constitutive aims of action (*A**). Those aims are inescapable, and hence ground reasons via Katsafanas' principle *SUCCESS* inescapably (cf. chapter 1, section 1). The norms that win out due to the reasons grounded by those norms and are compatible with *power* (and *agential activity*) are not just those two aims of action themselves, however. They are indirectly explained as endorsed via reasons stemming from the aims.

To what extent are the norms that Katsafanas explains really moral norms? He is, to some extent, a moral revisionist – unsurprisingly for a Nietzschean. But even *qua* moral revisionist, he tries to remain egalitarian, and he tries to justify his suggested norms even when they compete with other norms, so they are conventionally moral in the sense that they are defended in first-order ethical debates (Katsafanas, 2013, pp. 218-231). Hence, his norms count as moral enough to fit P_{PM} for present purposes.

However, I am very sceptical about Katsafanas' aims of action. Here, I focus on the aim of power. What does it mean to say that action aims at power? It is not fully clear what Katsafanas takes 'power' to be, though he summarizes his account by saying that Nietzsche's will to power is to be understood as the activity of overcoming resistance: 'to will power is to perpetually seek to encounter and overcome resistance in the pursuit of some end' (Katsafanas, 2013, p. 159). Resistance, in turn, should be understood as an impediment or challenge to the satisfaction of one's end (Katsafanas, 2013, p. 158).

Elaborating on this, we get:

> (*POWER*) An agent *A* has power (to some degree) over a resisting object *O* if *A* intentionally overcomes *O*, where (*i*) 'object' specifies a (token of an) entity, broadly construed, aspects of which provide resistance for the satisfaction of *A*'s end, (*ii*) 'intentionally overcomes' means that *A* intentionally succeeds in understanding, using, or becomes able to use *O* (to some degree), and (*iii*) *A* has more power over *O* the more *A* intentionally overcomes *O*.

For example, assume that Gabriel's (*A*'s) aim or end is to learn how to play the guitar (*O*). When learning how to play the guitar, Gabriel overcomes, and therefore has *POWER* over, aspects of playing the guitar such as musical scales and how to make use of the strings and the fretboard to produce the right notes (as per (*i*)). As Gabriel develops his skills, he intentionally overcomes the resistance offered by these aspects of the guitar. And Gabriel can master playing the guitar to various degrees (as per (*iii*)) – the more he understands music and develops his abilities to use the instrument, the more he overcomes the resistance it provides.

Plausibly, then, we can interpret 'willing power' as a higher-order aim, much like Velleman's self-knowledge (cf. chapter 1, section 1; section 3 above). Action tokens have particular ends, but as one pursues them, one also pursues the higher-order aim of power.

And that pursuit, Katsafanas thinks, is in itself active; we strive to encounter resistance to the satisfaction of our aims or ends, and we can do so to various degrees. Gabriel does not just try to learn how to play the guitar to be able to play it; he seeks to master it in some more complex way.

The higher-order aim of *POWER* is, Katsafanas thinks, present in every token of action. But, of course, such a controversial psychological position must be defended. Katsafanas' argument for thinking that *POWER* is an agential aim is based on a Nietzschean drive psychology. It goes like this:

> First, drives are motivational states that aim at their own expression, and take various objects merely as chance occasions for expression. Second, drive-motivated actions constitutively aim at encountering and overcoming resistance. Third, all human actions are drive-motivated. It follows that all human actions inescapably aim at encountering and overcoming resistance. (…) [A]ll human action manifests will to power. Power is a constitutive aim of action. (Katsafanas, 2013, p. 165)

Some clarificatory comments are necessary before I start discussing the argument. What is a 'drive'? Katsafanas attributes four main properties to them: '(*i*) they are dispositions that generate affective orientations; (*ii*) they admit an aim/object distinction; (*iii*) they dispose agents to seek their aims, rather than their objects; and (*iv*) they are constant' (Katsafanas, 2015, p. 165).

More specifically, by (*i*), drives induce representations of the world to be affectively charged. For example, assuming that hunger is a drive, food looks more appealing to an agent who is hungry than to one who is not. Furthermore, (*ii*) says that the aims drives have can be achieved by different objects. Different kinds of food can satisfy hunger. Moreover, (*iii*) drives aim not at their objects but at their own expression. An agent seeks a meal (object) to satisfy her hunger (aim), not the meal for its own sake when driven by hunger. And (*iv*) the drive to eat is constant, even though it can be temporarily satisfied. But it will recur, so a temporarily full agent will presumably seek out food to eat again in the future.

How do drives aim at overcoming resistance? They are aim-oriented rather than object-oriented (cf. (*iii*)). This means that they are directed towards being continuously acted upon over time, not just fulfilled at a single particular instance (in the way a desire might be). Hence, Katsafanas takes actions that are motivated by drives to be *process-directed*, not *goal-directed* (as they would have been if they were motivated by ordinary desires). Process-directed actions aim not at their own satisfaction, but at continuously succeeding at understanding, using or becoming able to use the objects necessary to satisfy their aims. Therefore, they aim at *POWER*, for having the aim of *POWER* involves exactly these properties. For example, the action of eating is presumably a way in which the hunger drive may reproduce itself. Because it nourishes the agent who performs it, it allows her to become hungry again in the future. And as it is part of the structure of the drive to find new objects to overcome; insofar as the agent engages in process-directed action, she aims at continuously overcoming resistance. All actions, Katsafanas thinks, are process-directed like this.

Can this argument be defended? I leave the first and second premises for now. As Katsafanas is aware, one does wonder whether the third premise is correct, however. Katsafanas indeed argues that all actions aim at the satisfaction of drives (Katsafanas, 2013, pp. 171-176). But here, I will argue that the evidence he uses to argue that actions are drive-motivated fails to support the positive conclusion, so apparently drive-based actions can be reinterpreted to fit into a more parsimonious, desire-based, framework.[16]

Katsafanas' main argument for the claim that we act based on drives rather than something else (e.g. desires) is an appeal to empirical evidence about the nature of satisfaction. There is some evidence which suggests that 'human beings are most satisfied when engaged in activities that provide them with challenges that are neither too easy nor beyond their capacities' (Katsafanas, 2013, p. 174). And that is supposed to support the view that engaging in process-directed actions may provide lasting satisfaction during the time one engages in them, while achieving ends by goal-directed actions does not. Goal-directed actions, Katsafanas thinks, never lead to lasting satisfaction. Hence, the evidence supports the view that actions aim at *POWER* and are process-directed.

However, these empirical findings are at best evidence for the weaker claim that there are *some* processes that provide lasting satisfaction, not that all actions work in this way. While we may be most satisfied by process-directed actions, it may still be the case that some actions are desire-based – I am sure that most people have direct first-personal experience of performing actions that have failed to lead to lasting satisfaction. Hence, Katsafanas has no argument against the view that some (other) actions do not lead to lasting satisfaction.

[16] He also spends some time replying to counterexamples to the third premise (Katsafanas, 2013, pp. 176-181). However, my discussion of the evidence leaves the counterexample discussion redundant; we can have a better explanation of what goes on in those cases without discussing drives if we can give a better interpretation of the data than one that supports drives.

Moreover, it is unclear why his empirical point could not be handled by saying that the seemingly process-based actions that supposedly may provide us with lasting satisfaction instead are goal-directed actions where we aim to satisfy desires that are suitably hard to satisfy. Aiming to satisfy such desires can be challenging in just the right way. So Katsafanas' positive case for drives does not establish that all actions stem from them; in fact, he has not conclusively established that *any* actions do (cf. Ferrero, 2015).

With this point in mind, an underlying problem for Katsafanas' view is that his drives seem to be unnecessary theoretical constructs. Many writers – including Katsafanas – admit that desires (at least sometimes) are motivationally forceful and may bring about actions. (And writers that do not tend to admit that other states or processes than desires provide motivational force that brings about actions, too.) But then, adding drives to agents' motivational structures adds an extra entity to action explanation that other theories need not involve. And as there is no evidence suggesting that we are better off by adding them to our ontologies, and it is unparsimonious to add them to action explanations, we have good reason to explain actions using desires (or something else) rather than drives. Hence, a purely desire-based view seems more explanatorily virtuous than his view.

How are we then to treat cases like hunger? Hunger does not look like any old desire; it has more of a drive-like structure. (It certainly fits Katsafanas' theory of drives well above.) However, presumably, one can supplement an account of actions motivated by desires with a take on paradigmatic cases of what Katsafanas calls drives where these are construed as dispositions to generate desires which *tend to* facilitate their own future desire-generation *without aiming* to do so. Dispositions and desires are not uncommon in action explanations, so we are not introducing extra entities with this move. Hence, a desire-based view still seems more parsimonious, and indeed more explanatorily powerful, than Katsafanas' drive-based one. So actions do not seem to aim at *POWER*.

(5) Smith

Michael Smith has recently published a series of papers (Smith, 2011; 2012*a*; 2013; 2015; 2017; cf. Smith, 2018 for elaboration) where he gives a constitutivist interpretation of his older practical reasons internalism (Smith, 1994, ch. 5; 1995). His argument has been presented in slightly different versions in these papers as well as in some recent talks, so it is somewhat hard to interpret. Hence, I will present a synthesized version of the argument that discloses its main moves – I do not aim to make it fit perfectly with any single one of Smith's versions, for it will be enough to present the general structure of his argument and then give a counterargument that applies to all versions.[17]

To provide a feeling for Smith's constitutivism, I want to immediately state that his main conclusion is that 'agents are morally obliged not to interfere with any agent's exercise of his rational capacities and they are also morally obliged to do what they can to make sure that agents have rational capacities to exercise' (Smith, 2011, p. 360). This is because 'every agent's fully rational counterpart has [desires to not interfere with any agent's exercise of his rational capacities and to do what they can to make sure that agents have rational capacities to exercise], [so by reasons internalism] every agent has the same reasons for action' (Smith, 2011, pp. 359-360).

I will call these desires 'desires to help and not interfere' (Smith, 2012*a*, p. 328), and I will use the terms '*manifestation*' and 'the *exercise* of' capacities interchangeably below. Terminology aside, however, Smith's main point is that agents' ideal counterparts' desires can provide us with reasons. This is because Smith is reinterpreting his older reasons internalism (Smith, 1994, ch. 5; Smith, 1995), according to which our reasons are based on what our idealized counterparts desire, so that it turns into a form of moral constitutivism. An idealized agent is fully rational, and Smith now argues that it is partially constitutive of being such an agent to have certain non-instrumental and dominant desires to help and not interfere, where their being 'dominant' means that they override potentially idiosyncratic ones which are not constitutive of ideal agency. Because having these desires is partially constitutive of being ideal, all fully idealized agent must have them. And because ideal agents' desires explain our reasons, we all have reasons to help and not interfere with the exercise of our own and others' rational capacities. These reasons explain our moral obligations to help and not interfere.

But let us start from the beginning. Why do we have such desires and reasons? Smith's argument has three main steps:

> (*STEP 1*) Assumptions about agency: agency is based on the Humean theory of motivation, is a goodness-fixing kind, and is ideal if fully coherent.
>
> (*STEP 2*) Full coherence requires ideal agents to have certain desires to help and not interfere. These desires generalize from

[17] In Leffler (2014), I discuss the 2011 and 2012*a* versions at much greater length. My interpretation here draws on that discussion, as well as on Bukoski (2016) and Lindeman (2019).

> helping and not interfering with our own exercise of our rational capacities to others' exercise of their rational capacities.
>
> (*STEP 3*) Desires to help and not interfere plus an ideal advisor theory of reasons generate reasons to help and not interfere. These desires explain moral reasons, which explain moral obligations.

I start with *STEP 1*. Agents have the capacity to act. For Smith, actions are events caused by instrumental desires, where instrumental desires are belief/desire-pairs that have been combined by a capacity for instrumental rationality (Smith, 2009; 2012*b*). Because actions involve beliefs and desires (as well as instrumental rationality), agency also has two sub-capacities: the capacity to believe truly, and the capacity to satisfy one's desires.

Moreover, Smith holds that agency is a goodness-fixing kind (Bukoski, 2016; Smith, 2017; cf. Thomson, 2008), so it sets the standard against which particular instances of it can be judged. As per the introduction of *GOODNESS-FIXING KIND* in chapter 1, section 1, toasters are such a kind. A good toaster toasts bread well, and a poor toaster toasts bread poorly, because toasting bread well is the function of toasters. A good agent, by contrast, is one who has and exercises her capacities to form true beliefs and satisfy her desires well – or, in other words, exercises the two sub-capacities of agency well. Furthermore, Smith adds that the ideal for the kind 'agency' is full coherence, or full rationality (Smith, 2012*a*).

STEP 2 of Smith's argument stems from the premise that an agent can fail to have or exercise these capacities perfectly. The capacities to believe truthfully and satisfy one's desires may conflict, and to avoid that, an ideal agent needs to have dominant desires to help and not interfere. The desires *not to interfere* are to be understood as desires not to interfere with the exercise of the agent's own rational capacities. Agents need them for the sake of avoiding present and future wishful thinking, for wishful thinking may make their psychologies incoherent, and hence less than ideal.[18]

'Wishful thinking' should be understood as believing or acting against one's best reasons for believing or acting because one's capacity for desire-satisfaction makes one form the wrong judgement or act on the wrong reasons. For example, a racist may form the belief that war refugees from another country are taking her job when arriving in her country – despite the fact that this is not the case – because she desires to find a reason

[18] In Smith (2011), he argues that we must have certain attitudes – of vigilance and reliance – to ourselves at different times, which in turn are constituted by these desires to avoid incoherence. I have omitted discussing these attitudes because they are an unnecessary middle-man in his main argument.

to think that the refugees ought to leave the country. Or she may try to dissuade them from entering her country, despite the fact that she has good moral reasons to let them stay. The agent's desire-satisfying and truth-believing capacities come into conflict here, so the agent is no longer ideally coherent. To avoid this kind of incoherence, the agent needs a coherence-inducing desire which dominates her other desires.

To be ideal, an agent must also have a desire *to help* her future self exercise her rational capacities. It might be the case that the agent will think wishfully in the future, because she may presently desire to believe or act against what her reasons support her doing in the future. A desire to help her future self would block such desires from generating incoherence, for again, lacking the right desires risks creating an incoherence-inducing conflict.

An ideal agent could also be incoherent now because she did not have a desire to help her future self in the past, because without such a desire, she could now be engaging in wishful thinking. Similarly, she cannot have allowed her past self to interfere with the deliberation she is currently conducting, because that might entail that she is now deliberating based on illusory premises. So an ideal agent must have had desires to help and not interfere all along.

Moreover, insofar as the ideal agents may lack or have impaired rational capacities, they 'must desire that (…) they do what they currently can to ensure that they acquire and maintain [their rational] capacities' (Smith, 2015, p. 191). This is why agents need desires rather than something else to stay coherent:

> An ideal agent's doing what's required to ensure the acquisition and maintenance of her epistemic and desiderative capacities requires that she has some desire that would motivate her to so act. Mere restrictions on the contents of her existing desires won't suffice. The addition of coherence-inducing desires to the psychology of an ideal agent is thus the only way to ensure that greater coherence can be induced in a uniform way across all of the cases in which greater coherence needs to be induced. (Smith, 2015, p. 191)

Finally, these desire to help and not interfere generalize to other agents' exercises of their rational capacities. Originally, Smith wrote that an ideal agent is entitled to rely on other (equally ideal) agents, so that they would not make her believe something illusory. Ideal agents must then generalize their desire not to make themselves believe something illusory to others on pain on making an arbitrary distinction between themselves and those they must rely on, since the other ideal agents have the same interests as they do. But ideal agents would not make such arbitrary distinctions (Smith, 2011).

Later, he has adduced other motivations for the generalization.[19] First, he has written that there is a number of symmetry-based arguments in the literature, e.g. Parfit's, that support his conclusion (Smith, 2012*a*).[20] Second, he has claimed that the generalization depends on two other symmetries. First, an agent's future possession and exercise of her rational capacities may depend on what she does now, and others' possession and exercise of their capacities similarly depend on what she does now. Second, just as an agent's present possession and exercise of her rational capacities may depend on what her past self has done, her exercise of her capacities also depends on what others have done (Smith, 2015). Because agents' capacities are intertwined in these ways, to treat like cases alike, '[ideal] agents must (…) be in the present such that all agents can possess and exercise their epistemic and desiderative capacities' (Smith, 2015, p. 192). Hence, our desires to help and not interfere generalize to covering others' capacities.

It does not matter which generalization consideration we go with here, however; what matters is that Smith defends generalization in one way or another. For regardless of which defence we prefer, the generalization move takes us closer to *STEP 3* of his argument. I have already explained the rudiments of how that step goes. An ideal agent's desires explain our reasons because that is how Smith's theory of reasons work. Our ideal counterparts all have desires to help and not interfere with the exercise of our own and others' rational capacities. These desires ground reasons for just that. And the reasons are universally prescriptive because everyone has them, and they are recognizably moral (cf. Smith, 1994, ch. 5; cf. chapter 1, section 2). Hence, they help explain our moral reasons to help and not interfere.[21] If moral reasons ground moral obligations, it follows that the moral reasons thus established also ground moral obligations – though, strictly speaking, this claim is not necessary for Smith's moral constitutivism. It is not clear why it matters that we have obligations rather than just moral reasons.

Regardless, Smith's view (*T*), then, involves a constitutivist defence of at least one part of morality. This kind of morality – reasons to help and not interfere with the exercise of one's own and others' rational capacities (P_{PM}) – can be explained by the desires we

[19] These may, though he does not say, be elaborations of his point in Smith (2011).

[20] In particular, Smith refers to Parfit (1984, pp. 140-141). Very roughly, Parfit's main point is that on theories where reasons are self-interested, there is an asymmetry: the theories treat an agent's own, but not others', future reasons as important.

[21] In recent, as of yet unpublished, work (Smith, 2018), he has started to argue the precise specification of these reasons depends on political negotiation in certain communities. Nothing hangs on this here.

must have if we are to be fully coherent, i.e. fully ideal agents. More specifically, our reasons are explained in terms of the desires (*A**) that our ideally rational, and therefore ideally constituted (*C*), counterparts would have, just in virtue of being fully coherent. And reasons are, everyone thinks, normative.

Just like on Korsgaard's view, then, *C* is here not to be understood in terms of aims implicit in action. But unlike Korsgaard, Smith understands *C* in terms of the constitutive features of *ideal* agency. So, to sum up Smith's view, the structural features of ideal agency give rise to our reasons, and some of these reasons are moral.

Morality vindicated? There are several problems with Smith's view (cf. Leffler, 2014). My biggest gripe with it is that I fail to see why fully ideal (or rational, or coherent) agents must have certain desires to be fully coherent rather than avoid losing their coherence for some other reason. Hence, I think Smith's move of adding these desires to ideal agents lacks support. And if that is right, we have no reason to think of ideal agents as having desires that ground reasons and obligations. This problem emerges on any of Smith's formulations of his arguments. It is of vital importance that the desires are added on any version of his view, for it is these desires that ground our reasons.

So why needn't we attribute the relevant desires to ideal agents? Above, I quoted Smith as saying that restrictions on the content of the ideal agent's desires would not suffice to guarantee that the agent remains coherent in the acquisition and maintenance of her capacities. Hence, we need desires.

But Smith's way of putting this point is based on a false dilemma. It need not be the case that the only two possibilities for how an agent can remain ideal are to either have restrictions on the content of her desires or to have Smith's desires to help and not interfere. There can be *other* restrictions on the ideal agent. The most obvious possibility is that an ideal agent may have to conform to a principle of rationality that prescribes that she does not become incoherent – indeed, this is closer to Smith's own (1994, ch. 5) coherence view of rationality than his newer suggestions. Then the ideal agent would not have a psychology that risks becoming incoherent while also counting as ideal, for if the restriction on her mental states is formulated in terms of rationality, then if the agent would fail to conform to it, she would *ipso facto* no longer count as rational, and hence no longer as ideal.

Alternatively, whether or not one takes coherence to be an assumption of rationality, perhaps one could supplement the ideal agent with some sort of Davidson-inspired holistic background coherence assumption according to which she can only form *more or less coherent* belief-desire sets, all of which are at least minimally coherent, rather than *incoherent* ones (cf. e.g. Davidson, 1980*a*). Such an assumption could immunize the ideal agent from severe capacity conflicts, such as those that Smith uses to get his argument going – the ideal agent would be psychologically unable to exercise her capacities that incoherently.

For yet another alternative, perhaps instead of saying that the agent must have such a rational capacity or coherence assumption, maybe coherence is just a condition of *idealization*. On this view, an agent who is ideal enough to explain reasons cannot be problematically incoherent, for then we might land with the wrong types of reasons in cases where our capacities conflict. In neither of these three cases are any desires to help and not interfere necessary.

Can Smith respond? Well, in Smith (2012*a*), he writes that agents must be able to exercise their capacities 'fully and robustly' (pp. 309-312). Agents must be ideal in a wide range of circumstances, including ones where their motivation differs from their present motivation. If so, adding desires to ensure coherence seems justified. But even if one wants to preserve this idea, it is not clear why one could not stipulate that the ideal agent must conform to, and exercise, capacities for rationality, Davidsonian coherence, or coherence from idealization fully and robustly instead.

In fact, there are even some positive reasons for making a change along these lines rather than adding extra desires. First, not adding extra desires to ideal agents seems more theoretically elegant than doing so. It is a theoretically clunky move, so a good motivation for doing it must be given, and one that fails to be better than alternative explanations is not.

Second, adding something else than desires to the ideal agent gives us a more general view. A good deliberating agent may well lack the psychological ability to become incoherent. She may, for example, be rationally sensitive in a way which makes her avoid it. Agents who are like that are possible, and we may even praise them for their character – they seem rationally stalwart. It follows that thinking of ideal coherence in terms of rational capacities rather than in terms of desires explains how such deliberating agents can be ideal too.

Might Smith reply in other ways? He could try to reformulate his argument. Doing so, he could say that the wishful thinking cases he adduces indicate that we have reason to think that there indeed are requirements of rationality that require us to have desires to help and not interfere. But the possibility of an agent without the need to have these desires, like the one in the paragraph above, shows that there is no need to posit certain kinds of desires.

A final possible reply is to appeal to the function of agency. In a recent reply to Bukoski (2016), Lindeman (2019) emphasizes Smith's appeal to the function of agency more than (I think) he does himself. Her thought is that any mental state that helps agents better realize their functions – to believe truthfully and realize their desires – is *ipso facto* justified to attribute to the agents. Hence, we are justified in attributing the relevant desires to their ideal counterparts.

I am not sure how this would improve on my alternative suggestions, however. At least the extra rational capacity (which is not desire-based) and the idealization suggestions should be able to do exactly the same work as extra desires do to preserve the agent's functionality across different situations. This matter is more complicated when it comes to the Davidsonian holist suggestion, admittedly, because if the function of agency involves full coherence for the purpose of exercising its sub-functions, the Davidsonian suggestion would still allow for some local incoherencies at times. The desires would do more to establish coherence than that. But at least two of my alternative suggestions still suffice to show that Smith has yet to establish why ideal agency requires desires to be fully coherent.

It seems, then, that it is not so easy to get constitutive desires out of Smith's general framework. I conclude that a constitutivist argument for moral conclusions should not be based on the view that certain desires are needed to solve wishful thinking problems.

(6) Conclusion

In this chapter, I have shown how many constitutivists try to explain moral normativity, and that when they attempt to do so, they make implausible theoretical assumptions. I have, therefore, argued that all the leading constitutivists in the literature have problems with their central background assumptions about action. More specifically, Korsgaard cannot explain why we need universalized maxims (and relies on implausible Kantian assumptions), Velleman cannot explain intentions because self-knowledge and control seem to come apart, Katsafanas cannot explain why we need to aim at overcoming resistance, and Smith cannot explain why it is constitutive of ideal agents to have desires that ensure coherence.

The problems here give rise to a more general hypothesis about how the constitutivist project in the literature looks when it comes to grappling with the problem of substantive assumptions. Constitutivists fail on the first horn of the dilemma, making too strong assumptions for their views to be plausible. Even worse, we can see that they do so because their views suffer from the problem of adequacy. The constitutivists' assumptions seem inadequate, and they do so because they seem false.

3. The Shmagency Objection

I have just argued that the leading constitutivists in the literature suffer from the problems of substantive assumptions and adequacy. Fine. But what about the arguably most famous problem for the view? This problem is known as the agency-shmagency problem, or – as I call it – the shmagency objection. This problem turns on various interpretations of the inescapability of agency. For while constitutivists disagree about which norms we are required to follow, most have also argued that the constitutive features are *inescapable*. Inescapability plays a vital role in the explanation of the normativity of the normative phenomena constitutivists attempt to explain, as well as in warding off objections to their views.

But we may question how inescapable the norm-explaining features are. If we do not instantiate the constitutive features that explain norms, it seems like we can avoid the norms they are supposed to explain. Someone who is a shmagent – very much like an agent, but without instantiating the norm-explaining features – is very similar to an agent, but, because the shmagent lacks the norm-explaining features, does not seem to be subject to the norms nor such that they have force for her (Enoch, 2006; 2011*b*). Hence, it seems like constitutivism is unable to explain norms for shmagents.

Here, I aim to show that, despite many constitutivist responses, new versions of the problem appear for most forms of constitutivism. In particular, it remains a deep problem for all the leading constitutivists I discussed in chapter 2. Hence, finding better solutions to the problem becomes necessary for any viable form of moral constitutivism.

To show this, in section 1, I present the original shmagency objection. In section 2, I show how the standard reply to the objection – that the shmagent is self-defeating because agency is descriptively inescapable – seems defensible, despite several arguments to the contrary. But then, in section 3, I extend the shmagency objection by arguing that shmagents can be sophisticated enough to have practical reasons while standing outside agency. This resuscitates the problem. In section 4, I explain how sophisticated shmagency remains a problem for some other recent constitutivist attempts to avoid the shmagency objection.

In section 5, I introduce another major line of response to the shmagency objection, according to which constitutivism is defended by appeal to constitutive features that are normatively inescapable instead. I call this view *partial constitutivism*. Because partial constitutivists think agency is normatively inescapable, it does not matter for their purposes if actual agents sometimes fail to be agents. But in section 6, I argue that partial constitutivism suffers from a second new version of the shmagency objection because they leave the norms that they are supposed to explain underdetermined. I conclude in section 7 by outlining how the new shmagency worries give rise to two *desiderata* for any constitutivist view.

(1) Enoch's Argument

The paradigmatic formulation of the shmagency objection comes from Enoch (2006). His basic point has often been set up using an example. Imagine that you are playing chess. There are certain rules (and maybe aims) constitutive of doing so; if you do not abide by them, you seem to be playing something else. Call this other game *shmess*. Why should you stick by the rules (or aims) of chess – rather than *shmess* – when you are deciding which game to play? A reason seems needed.

By analogy, Enoch thinks, it is unclear why we should care about what is constitutive of action or agency. We can always ask 'So what?' and demand a reason for why we should be agents rather than shmagents – something very much like agents, but not quite like agents. Or, to put the point more poignantly, we can ask the *shmagency question*: 'Why should I be an agent rather than a shmagent?'

The question is meant to illustrate that we can avoid being agents by being shmagents instead. We can, so to speak, shirk from the normative phenomena that agency is supposed to explain for us. For if we are shmagents rather than agents, we can have all the features that we would take to be constitutive of agency – or even otherwise associated with it – except those that explain the norms that hold for us.

But as constitutivists attempt to show how normative phenomena hold and are normative for agents in virtue of properties of constitutive features of agency (including its inescapability), then if we can be shmagents, it seems like their story does not get off the ground. If shmagency is an open option for us, then leading constitutivists have yet to show how normative phenomena are normative, for they have not explained why we are subject to them or why they have force for us.[1] Hence, when I mention the shmagency question below, I take its main point to be equivalent to suggesting that agency is not

[1] At least, this is the standard interpretation of the objection. Alternative interpretations are discussed by Katsafanas (2013, ch. 2), Paakkunainen (2018), Rosati (2016), and Smith (2015), usually along with this one.

comprehensive enough to show how the normative phenomena it is supposed to explain apply to, or have normative force for, agents.[2]

More formally, here is the problem:

> (1) If constitutivism is true, the constitutive features of agency in virtue of which some relevant set of normative phenomena are normative for us must be (descriptively) inescapable.
> (2) We can (descriptively) escape instantiating the constitutive features of agency in virtue of which some relevant set of normative phenomena are normative for us.
> ---
> (C) Constitutivism is false.

The core reasoning here is already present in the description of the argument above. The thought behind (1) is that if we can escape the constitutive features of agency in virtue of which norms are normative for us, then we do not have a story about the phenomena these features are supposed to explain. The thought behind (2) is that we indeed can avoid instantiating the properties of agency that do this, for we can be shmagents. (Or, equivalently, we can ask the shmagency question.) (C) follows immediately.

Some clarifications are, however, needed before I proceed to discuss the argument. First, I have written 'the constitutive features of agency in virtue of which some relevant set of normative phenomena are normative for us.' The features of agency in question should be interpreted as substantive, and the relevant normative phenomena are normatively forceful reasons for action, some of which are moral reasons.[3] Whatever else Korsgaard, Velleman, Katsafanas and Smith talk about in their frameworks, they all appeal to substantive conceptions of agency and aim to explain normatively forceful practical reasons, some of which are moral. Hence, the formulation captures all their views.[4]

[2] This core point can be extended further with other assumptions – e.g. if we need reasons to be agents, then those reasons may need to be external to agency, so constitutivism cannot explain all reasons.

[3] What does 'normatively forceful reasons for action' mean? In Leffler (2019), I wrote that 'a practical reason, *pro tanto* or overall, for an agent *A* to φ is normatively forceful iff the reason cannot legitimately be ignored because *A* arbitrarily desires or wants something else than to φ' (Leffler, 2019, p. 124). This leaves many theoretical options open, however, and is fundamentally intended to be a way to capture the views of Korsgaard, Velleman, Katsafanas and Smith.

[4] Of course, if the reader thinks that her favourite form of constitutivism suffers from the shmagency objection even though it is some other form of constitutivism, she should feel free to reinterpret the rest of my discussion in that way.

With this point in mind, when I talk about 'constitutivists' in this chapter, I mean to talk about constitutivists with those commitments. It is possible that other constitutivists, especially with weaker commitments, will be able to handle my arguments. In fact, to precede the discussion in chapters 7-8 below, I shall propose just such a view myself. But how that works will emerge later.

Second, the notion of inescapability in the shmagency objection is fairly complex. The standard interpretation of inescapability is that it is some descriptive form of necessity, not normative necessity. In particular, I shall start off by assuming that the inescapability involved here is descriptive. However, this assumption will be tweaked below; in sections 5 and on, I will discuss normative inescapability, according to which it is normatively desirable to be the kind of agents that can explain norms.

Instead, for now, assume that the kind of inescapability that is involved in the shmagency argument is *dialectical inescapability*.[5] Dialectical inescapability is a descriptive form of inescapability, for it is something that an agent has, rather than one that she ought to have. Luca Ferrero characterizes it as 'the inescapability of rational agency in the sense of the closure of this agency under the exercise of its distinctive operation' (Ferrero, 2018, p. 128). What is inescapable is the agency that an agent already has, and agency is inescapable because it is self-defeating to attempt to escape agency, because acting so as to escape it involves exercising one's agency. It is this form of inescapability that most writers have had in mind when trying to ward off the shmagency objection.

(2) Inescapability and Self-Defeat

The most common reply to the shmagency objection, then, is to deny the argument for premise (2). We cannot, it is claimed, properly ask the shmagency question. This response comes from a dilemma based on a distinction between an internal and an external way to ask it. The question is internal if it is asked by someone who already is an agent, or external if it is asked by someone who is not. The internal question is largely unproblematic, for it is a normative question whether an agent should be an agent. Maybe one should not, or not always, be an agent, but at least constitutivists can try to give reasons for or vindications of why one should be an agent as soon as one has come this far. And as long

[5] However, as Ferrero (2018) points out, this kind of inescapability need *not* be the kind that many constitutivists think explains normativity. What positively may explain normativity is orthogonal to the present discussion (cf. chapter 5, section 6, below for discussion).

as agents remain agents, constitutivists can provide whatever positive explanation of practical reasons they want.

However, according to the standard reply, the external question does not arise. The most important reason for thinking that it does not is that anyone asking the question already is an agent, so it is self-defeating to ask it. Asking the external shmagency question is still an action, and hence subject to the norms explained by agency. Hence, agency is (dialectically) inescapable.

One version of this response, paradigmatically formulated by Ferrero (2009), has generated most of the ensuing discussion (cf. Enoch, 2011*b*; Ferrero, 2018; Katsafanas, 2013, ch. 2; Korsgaard, 2009, ch. 1; Rosati, 2016; Velleman, 2009, ch. 5). I will start off by defending this argument, and hence constitutivism, against some recent responses. However, in the next section, I shall point out a deeper shmagency problem, hence criticizing premise (2) anyway.

According to Ferrero's response, then, agency is dialectically inescapable in virtue of two properties (Ferrero, 2009, pp. 308-309). First, agency is *the enterprise of the largest jurisdiction*, so all actions fit within its scope. Playing chess and playing shmess are both actions, but a shmagent cannot act in the same sense as an agent. Second, agency is *closed under reflection*, meaning that reflecting on how to escape agency, let alone actively trying to do so, still counts as acting. It is still logically possible to opt out of agency, e.g. by committing suicide. But once one is an agent, one cannot deliberately avoid being an agent without exercising one's agency.

The key argument, then, is that because agency is the enterprise of the largest jurisdiction, and one cannot opt out of it in the same way that one may decide to play shmess rather than chess, there is no alternative to it once one is in the game. One cannot deliberately leave for something else without exercising it. Hence, it is self-defeating to ask the shmagency question for an agent.

I will proceed by presenting three points that can be construed as replies to the charge that escaping agency, and hence becoming a shmagent, is self-defeating. First, Enoch provides two such considerations. Responding to an interpretation of Velleman according to which Velleman considers it constitutive of agency to *care* about one's constitutive aim – so caring about it is inescapable, and this explains why we are subject to norms – Enoch writes:

> [W]hat we are up against here is an *especially* problematic instance of [a naturalistic fallacy]. (…) I want to concede that agency is

> indeed naturally inescapable for us. But I also want to note (…) that such inescapability does not matter in our context (…). For the move from 'You inescapably φ' to 'You should φ' is no better – not even the tiniest little bit – than the move from 'You actually φ' to 'You should φ.' (Enoch, 2011*b*, p. 216, his italics; cf. p. 211)

The objection here is that constitutivism suffers from a version of the naturalistic fallacy. But the objection shifts the topic. It does not seem to have much to do with dialectical inescapability. As Ferrero (2018) points out, dialectical inescapability need not by itself be used to explain why norms have force for agents. Because it is not the property to which constitutivists appeal to explain why norms have normative force (and hence are such that one *should* follow them), dialectical inescapability is not subject to a naturalistic fallacy. Dialectical inescapability only shows why one cannot avoid agency once one is an agent, and hence why norms apply to agents. But explaining why norms have force is not the same thing as explaining why norms apply. Norms of etiquette apply to agents, but whether or not they have force is another question (cf. Foot, 1972).

Instead, the positive explanations of normative force constitutivists provide tends to depend on some other inescapability-related property, such as Korsgaard's plight inescapability (Korsgaard, 2009, pp. 1-2; cf. chapter 2, section 2; chapter 5, section 6). On my interpretation, according to plight inescapability, we cannot avoid being subject to norms because we keep facing new choice situations in which we must act, so we must continuously face the demands of agency, *and* this is part of the explanation of dialectical inescapability. But it is the former conjunct which may help explaining why norms have force. The naturalistic fallacy charge is aimed at the positive explanation that a property like plight inescapability might provide, as is the talk about a move from 'You actually φ' to 'You should φ.'

Having said that, it could be argued that this response of Enoch's still generates a problem for constitutivists, as it shows that they have to say more to explain normativity. But that is a point that constitutivists happily may concede and then go on to do so (cf. chapter 2; 5; 6; 8). Naturalistic fallacies seem beside the point at the present stage of the shmagency dialectic; constitutivists are allowed to say that one cannot avoid agency and then supplement their explanation of normative phenomena with any story about why norms have force that they want.

Enoch's second response is that constitutivists reify the sceptic into an actual character that they try to convince. They try to show that the potential shmagent cannot escape its agency predicament. That means that they do not deal with the conceptual problem that the shmagency objection stems from. As he puts it, the sceptic 'is not (...) an actual character, with a position to defend, [but] the embodiment of a problem *we* face, because of *our* commitments' (Enoch, 2011*a*, p. 219, his italics).

The shmagency objection should instead be understood as a problem for our concept of agency. The challenge is that constitutivism does not show why we should be agents *even* if the shmagent is self-defeating because she asks the shmagency question. The self-defeat response would, in a way, be an *ad hominem* charge of hypocrisy against the shmagent. But such hypocrisy is irrelevant – hypocrisy does not imply that our concept of agency is such that there is no question to ask about why one should be an agent. It only shows that the hypocrite is in a place where asking the shmagency question becomes hypocritical because she already is committed to agency. But whether she should be an agent is what is at issue.

Ferrero has replied to this point by conceding that there is a sense in which he treats the shmagent as an actual character, but also claims that this does not matter (Ferrero, 2018, p. 131). Here Ferrero relies, again, on the distinction between internal and external questions. The shmagent occupies a position external to agency and asks whether it should become an agent, but that position can be shown to be self-defeating (by the argument above). This leaves the internal question – why an agent should care about being an agent, rather than a shmagent – open. But the reply to the internal question is distinct from the dialectical inescapability of agency, which can defuse the external question. Again, constitutivists can respond to the internal question in any way they want. It is enough for them to avoid the external one.

A third reply to the inescapability worry comes from Tiffany (2011). Tiffany accepts Ferrero's point that the external question is self-defeating. However, he also holds that the kind of agency one cannot opt out of is too minimal to explain strong normative standards, such as those that forceful reasons like moral ones can provide us with. Hence, Tiffany thinks, some form of constitutivism may be true about some weak norms, but not stronger norms.

The underlying reason for this is that he believes that constitutivists equivocate on the nature (or, possibly, concept) of agency. According to Tiffany, just because we cannot opt out of some weak form of agency, it does not follow that agents cannot opt out of substantive constitutivist-style agency that might explain norms. Maybe a minimal agent can be an agent in some sense – for example, she might be able to act for reasons. But the constitutivists I am discussing here start off from substantive theories about agency that involve more than minimal agency. For example, Kantian constitutivists like Korsgaard think that agents are committed to *CI* (cf. chapter 2, section 2). The minimal agent need not be committed to anything that strong.

However, Tiffany's equivocation response goes by too quickly. It seems to beg the question. Constitutivists often attempt to explain reasons in terms of some features of (presumably intentional) agency as such (cf. Katsafanas, 2013, pp. 37-46). This means that there is no weaker version of agency out there – or, at least, there might only be one type of relevant agency, whereas other types of agency (e.g. animal agency) are extremely different and hence need not have the same normative commitments (cf. Korsgaard, 2009, ch. 3-7; 2018). So constitutivist theories of agency differ from minimalist theories not by taking there to be different standards of agency for different (relevant) agents, but by claiming that agency involves much more than some weak standard like the ability to act for reasons from the start. If the constitutivists are right, it follows that every agent (or every normatively relevant agent) is committed to everything that agency involves.

But might one not think that there are several forms of normatively relevant agency from the start, like Tiffany and others appear to do (cf. Lavin, 2017)? Why would agency be unified so as to generate the same normative reasons for all? This option is, admittedly, theoretically open. But absent an argument in its favour, it still seems question-begging. Constitutivists can answer: why should we believe that there is more than one kind of (normatively relevant) agency? More would have to be said to give constitutivists a reason to go with a disunified account.

(3) Normativity for Shmagents

I have just presented three lines of defence of the inescapability reply to the shmagency objection. There still seems to be a sense in which at least the standard kind of agency remains dialectically inescapable, and the sceptic therefore self-defeating. However, I shall now argue that premise (2) remains defensible. Even though the original shmagency question can be avoided, the objection can be extended in a way that resuscitates the original problem.

How so? The final response to the shmagency objection that I discussed and criticized was Tiffany's equivocation response. Even though it begs the question, there is still something to his point that different ways of being might generate different normative results. We may well accept the constitutivist response to the shmagency question and make a deeper point that threatens premise (2).

This is because insofar as we want to explain strong norms, such as those of moral or otherwise forceful normative reasons, we need to know more than who is an agent. Regardless of what agency involves, as long as the constitutive feature(s) used to explain morality is relatively complex – and all constitutivist views under discussion are in agreement here – constitutivists will have trouble giving a good enough explanation of the normative reasons that hold for many creatures that appear to have them. Some of them can stand outside agency and ask the shmagency question. Accordingly, constitutivism does not seem to provide a good explanation of reasons. It cannot explain the reasons of some creatures who seem to have them, and hence lacks explanatory power even when construed as a theory about some subset of the reasons there are.

In particular, constitutivists cannot explain reasons for what I will call *sophisticated shmagents*. Sophisticated shmagents appear to have reasons and stand outside agency, so such shmagents can ask the external shmagency question. This vindicates (2) in the shmagency argument. (We can call the fact that they appear to have reasons, or at least something reasons-like, *the problem of normativity for shmagents*.)

But who are sophisticated shmagents? I stipulate that they are shmagents who are intelligent, knowledgeable, and perform what looks a lot like actions for what looks a lot like reasons – and, I shall argue, what well may be reasons.[6] They are also capable of (what looks like) deliberation, reflecting about what to do, and are able to prefer different actions to different extents. Hence, they seem like *prima facie* good candidates for participating in ordinary normative practices, such as that of giving reasons for their actions when asked why they are doing what they do.

But sophisticated shmagents cannot act and are not agents according to constitutivists. This is because they lack at least one – possibly all – of the constitutive features of agency that constitutivists also use to explain reasons. Since constitutive

[6] If one wants to use the word 'reasons' conservatively, one may call what sophisticated shmagents have 'shmeasons' – but they still seem to have exactly the same kind of role and force as reasons do for agents, so the 'shmeasons' still seem equivalent to reasons. Hence, if we think more deeply about who the creatures that lack the reasons-explaining features are, it does not seem like they do not have reasons – instead, they seem to function surprisingly much like agents (who have reasons), which indicates that we have good reason to think that they have reasons.

In fact, moreover, we have no pre-theoretical reason to think that what they have should not be explained in the same way as the reasons agents have. And it would be fallacious to think that because constitutivists can explain reasons for agents by appealing to the constitutive features of agency, there are not some reasons for shmagents that they cannot explain. Even though constitutivists can explain one part of the normative sphere (for agents), they may well be even more to it.

features are necessary features, without them the sophisticated shmagents fail to qualify as agents.

We can concede to constitutivists that agency should be understood in their preferred ways; in fact, as I argued *contra* Tiffany, we should do so, or else we beg the question against most constitutivists. But even with that concession made, there remains a conceptual and normative space where sophisticated shmagents can operate. And if they can do so, a problem re-emerges for constitutivism. Sophisticated shmagents can ask the external question about whether they should be agents, i.e., they can reason practically about whether or not they should be agents, since they have what appears to be reasons. However, being shmagents, they still stand outside agency – in other words, they are external to agency. So it seems like they can ask the external shmagency question.[7]

They can even do so independently of the dialectical inescapability of agency. To rehearse the last section: my responses to objections to Ferrero's argument were (*i*) that the naturalistic fallacy point does not matter because the fallacy has little to do with dialectical inescapability, (*ii*) that it does not seem to matter that constitutivists reify the shmagent because they can still defuse the external question, and (*iii*) that because constitutivists think agency involves a lot from the start, it begs the question to hold that only minimal forms of agency are inescapable.

Yet none of these responses indicate that there cannot be sophisticated shmagents. The responses can be avoided as follows: (*i*) the explanation of reasons in terms of agency is neither here nor there if we can escape agency, which sophisticated shmagents can. (*ii*) Because sophisticated shmagents stand outside agency from the start, they can ask the external shmagency question. (*iii*) Sophisticated agents are shmagents, *ex hypothesi*, so they have little to do with what constitutivists take agency to involve.

With these points in mind, I shall illustrate how shmagents appear to have reasons while standing outside agency, vindicating premise (2) in the argument, by discussing Korsgaard's theory of agency. Assume that she is right. As per chapter 2, section 2, she thinks that acting involves acting on maxims, these need to be universalized, and universalization must proceed in line with *CI*. If it does not, we are mere heaps, not agents.

[7] I suspect that sophisticated shmagency is a problem here because philosophers have thought of shmagency as an offshoot of agency to too high an extent. Shmagency is usually taken to be agency minus the normativity-explaining feature that agency purportedly has (and possibly minus something else, but only little else). Hence, one might become a shmagent if one loses some agency-constituting feature. But shmagents can be extremely cognitively and maybe even normatively sophisticated, in the manners just described, while standing outside agency. They need not be 'agents minus'.

But assume simultaneously that we have some sophisticated shmagents. Call them the Martians. A traditional Humean belief-desire theory is true about how the Martians behave or otherwise interact with their environment (rather than act, since only agents can act), intentionally or not, and their ordinary behaviour or interaction usually stems from belief-desire combinations of mental states. These are not in any way regulated by *CI*. There is no need to appeal to maxims, universalization, or being in line with *CI* to explain their behaviour; such features, which Korsgaard takes to explain why we are bound by *CI*, are in no way part of their psychologies. Hence, they lack the norm-explaining features that she thinks are constitutive of agency.

Or consider some other sophisticated shmagents – the Saturnians – whose behaviour or interaction stems from besires, i.e. mental states that both represent the world and push them to behave in certain ways. Again, they lack the features that might seem to bind us to *CI*. Examples of creatures with different kinds of psychological setups can be multiplied pretty much indefinitely here; they all lack the features that are constitutive of agency and constitutivists take to capture reasons. I focus on these two, however, as they exemplify psychologies that philosophers often have thought explain action.

The Martians and Saturnians fail to qualify as agents on the theories of agency that constitutivists like Korsgaard hold. There is, *ex hypothesi*, no way that their 'actions' have typical constitutivist structures or aims. Again, maxims, universalization, or being in line with *CI* have nothing to do with the explanation of Martian or Saturnian behaviour. It follows that they do not aim at following norms such as *CI* (in any relevant way, at least), and hence do not have reasons on Korsgaard's view.

Yet they still appear to have reasons. They are sophisticated and are therefore, pre-theoretically, on par with at least humans insofar as reason-possession goes. I have already assumed that sophisticated shmagents have all kinds of properties that indicate that they have reasons: they are intelligent, knowledgeable, perform what looks a lot like actions for what looks a lot like reasons, are capable of (what seems to be) deliberation and reflecting on what they do, and are able to prefer different behaviours. And they seem to be *prima facie* good candidates for participating in normal normative practices.

Moreover, here is a number of things we take to be true about reasons. First, (*i*) they are facts counting in favour of something (for someone), so they would have to be reasons for someone. Furthermore, (*ii*) they appear to be normatively forceful (at least for those who have them), (*iii*) they depend (e.g. supervene on, or are grounded in) natural facts, (*iv*) they come in varying strengths (or weights), (*v*) they can contribute to generating all-things-considered reasons, and (*vi*) they have impact on deliberation.

All the properties that reasons are supposed to have seem possible to instantiate without having a Korsgaard-style constitution. For example: let a Martian deliberate (or deliberate*, if you want to reserve the word 'deliberation' for a kind of action that constitutivist-style agents perform). Let it also deliberate using facts; its desires might be backgrounded (Pettit and Smith, 1990). From its perspective, when deliberating, it represents facts – which may or may not seem desired.

Using representations of these facts, the Martian judges which ones count in favour of what to do (*i*). It is these facts that appear, to the Martian, to be relevant to determine what it is to do by favouring different outcomes (*ii*).[8] Moreover, the facts that it takes into consideration seem to stem from natural properties, such as whether or not something is pleasant or painful for the Martian (*iii*).

Now, the Martian thinks these facts matter to different degrees (*iv*), but weighs them up, and tries to reach a conclusion about what to do based on what it most strongly favours. Because it is knowledgeable and intelligent, it can do this to quite a significant extent. What appears to be reasons therefore comes in different strengths (*v*), and these seeming reasons are part of generating what looks like an all-things-considered reason. The Martian, then, seems to be deliberating with reasons (*vi*). And all this could be said about the Saturnians as well.

Again, it certainly seems like the Martians or Saturnians have reasons, or at least something that plays the role of reasons. Because these creatures stand outside constitutivist-style agency, there is a perspective from which it makes sense to ask the external question about whether they should be agents or shmagents. So we can preserve premise (2) in the original argument. So the shmagency objection stands.

Unexpectedly, there are some worries about this defence of the shmagency objection. First, it might be claimed that it does not matter that there are shmagents who occupy a position outside agency from which they may ask the shmagency question. Maybe what matters is whether agency is inescapable for creatures *like us*, who already are agents. If it is, for example, psychologically impossible for us to have belief-desire psychologies or besires because we are the kinds of creatures who have Korsgaard-style

[8] It might be thought that the 'normative force' here is fairly weak – the only force I assumed is that the reasons seem relevant for determining what to do to the Martian. But the reader is free to slot in many possible theories here (as long as they are compatible with some facts seeming that way to the Martian).

constitutions, and these cannot be altered, becoming shmagents is not a live possibility for us.[9]

It is possible that we might not be able to become shmagents. I have defended the inescapability of agency for agents in section 2, and then tried to argue that the real shmagency problem comes from creatures like the Martians or Saturnians. Such shmagents may never have been agents in the first place.

But because they still seem to have reasons, or something very much like reasons, *that* normative phenomenon should be handled in the same way as reasons for ordinary agents (cf. footnote 6 above). We should want a general theory of (what looks like) reasons. But constitutivism seems ill-suited to make sense of a phenomenon like reasons if it is limited to agents' reasons and not the reasons of sophisticated shmagents – it only seems able to make sense of a subset of our observations of the reasons there are. Constitutivism does not seem very explanatorily powerful if it only can do that, for then it cannot handle all the reasons there seems to be. Hence, it is likely false – even about reasons for ordinary human agents.

A second possible response is to say that the kind of reasons that shmagents have are somehow different from, and probably of less normative interest than, those that constitutivist-style agents have. Or, similarly, one might think that what I have called shmagents are agents of another kind than standard constitutivist agents, and then argue that one should explain their reasons in different ways.

Versions of this point are already made in the constitutivist literature. Most obviously, Lavin (2017) thinks there can be different kinds of agents who have different kinds of norms applying to them in virtue of having different kinds of constitutions. Similarly, Korsgaard (2009, ch. 3-7; 2018) thinks that while humans indeed act in accordance with universalized principles – and our fundamental principle is *CI* – for animals, instincts do the work of our principles. By distinguishing between different kinds of principles, we can make sense of different kinds of agency, and, possibly, different kinds of reasons that stem from different sources.

But appealing to different kinds of reasons seems disingenuous insofar as we are talking about agents and shmagents instead of when we contrast, for example, human beings and animals. We can easily make further assumptions about the Martians and Saturnians that explain why it seems like they have reasons in the ordinary sense of the word. When I characterized them, I stipulated that they have all kinds of properties that

[9] Thanks to an anonymous reviewer of Leffler (2019) for raising this point.

make them seem to have reasons, and I have also showed how what looks like reasons might feature in their phenomenology. Examples of properties needed to have reasons in the ordinary sense of the word can be multiplied, since the Martians and Saturnians are creatures that we construct. They can always be made sophisticated enough to seem to have reasons in the ordinary sense.

(4) Other Reasons to Dismiss the External Question

I have now motivated the shmagency worry again. Sophisticated shmagents seem to have reasons and stand outside agency, so they can ask the external shmagency question. However, some other motivations than the self-defeat argument from section 2 have also been proposed for explaining why the external question fails to make sense. If either of these is right, premise (2) is defended. I shall first discuss a semantic response, and then a metaphysical one.[10] I shall argue that these responses, too, fail due to considerations that have appeared in the discussion in the previous section. The responses attempt to show that it is impossible to ask the external shmagency question, but sophisticated shmagents can do so.

First, there is a second strand of argument in Velleman's response to the shmagency objection.[11] The idea is that the external question – asking 'Should I be an agent or a shmagent?' from the perspective of a non-agent – does not make sense because it is semantically defective. It looks analogous to 'Is a tree taller?' without specifying what the tree might be taller than. If so, it is conceptually impossible to ask the shmagency question, and so the possibility that we might end up outside agency is no challenge to constitutivism.

However, as Enoch (2011*b*) points out in his response to the argument, shmagency-style questions do not seem defective. Asking whether one should be an agent

[10] Some responses can be treated more quickly. O'Hagan (2014) seems to endorse both response types, but she is not able to face Tiffany's challenge from section 3. She tries to defend constitutivism by arguing that shmagents must deny a minimal norm of reasons-responsiveness, but it does not follow from that norm that constitutivists have enough to say about agency to be able to explain more or otherwise forceful practical reasons.

Furthermore, Rosati (2016) argues that the difference between agents and shmagents is greater than what Enoch has assumed, but once we see that, we realize how much more valuable agency is to us. So agency matters because it is valuable. But, obviously, that requires her to have a take on values that is independent of what we can squeeze out of agency if it is to answer the sceptic. I discuss such responses in sections 5 and 6, below.

[11] Or, rather: Enoch (2011*b*) attributes it to Velleman. I am not sure about whether Velleman himself endorses it.

or something like it, or whether one has reason to be an agent, seems perfectly intelligible. So, *prima facie*, shmagency-style questions do not seem defective.

Yet seemings can be erroneous, so maybe we need a deeper reason to think that they are correct in this case. One such reason stems from the point that the intelligibility of the external question does not stand or fall with the possibility of shmagency for creatures who already are agents. If the external question is unintelligible, it would make no sense for sophisticated shmagents, standing outside agency but having reasons, to ask the shmagency question. But whether sophisticated shmagents should become agents or not clearly matters for them – assuming some constitutivist view is correct about the normative commitments and implications of agency, they would be subject to different norms if they were to become agents, which no doubt matters from their perspectives. This would not have made sense if the shmagency question had been conceptually confused. Hence, the external question does not seem semantically defective.

Another attempt to motivate the failure of the external interpretation of the shmagency question comes from Silverstein (2015). He thinks that it makes sense to ask it, but that it is ambiguous between its internal and external senses. The internal question makes sense, but the external one does not. A shmagent – who is not an agent – would be asking for reasons for actions though she has none, but anyone asking the question is already an agent. So the external sense of the question begs the question against the constitutivist picture of normativity, according to which it is agency that explains why something is a reason:

> It is tempting to interpret the shmagent's question as one about reasons for action: Do I have any reason to become an agent rather than a shmagent? But that cannot be right, for a shmagent is not in a position to perform actions. Only agents can act, and so only agents can be in the market for reasons for action. (Silverstein, 2015, p. 1138)

However, this answer is unsatisfactory due to the problem of normativity for shmagents. Sophisticated shmagents are still in the market for reasons (or something reasons-like) for action (or for something action-like), and the external question certainly seems intelligible for an intelligent being who does not count as an agent according to the strong constitutivist theories of agency currently under discussion. So again, the external shmagency question remains a live possibility.

(5) Partial Constitutivism

I have now argued that there are shmagents who plausibly have normative reasons, and hence defended premise (2). Can constitutivists respond to the shmagency objection in a better way? A number of authors have recently defended views according to which we – most directly – should be agents normatively rather than descriptively.

For example, Michael Bratman (2016) has argued that norms constitutive of planning agency are justified in virtue of their value for our self-governance. And though she does not commit herself to constitutivism, Caroline Arruda (2017) has argued that the reason we ought to be full-fledged agents is not, as constitutivists have argued, that full-fledged agency is inescapable, but because it allows us to pursue other valuable projects that require exercising full-fledged agency. It seems easy enough to turn her point into an argument for a form of constitutivism saying that we ought to endorse norms constitutive of full-fledged agency because of their general value for us.

Most importantly, however, Michael Smith's version of constitutivism (cf. chapter 2, section 5) diverges from older forms of constitutivism in interesting ways. Because his view is the most developed theory in print according to which a deeper norm allows us to explain reasons in a constitutivist way, I shall use it to exemplify the second response strategy to the shmagency objection.[12]

We can call all the views I just have mentioned versions of *partial* constitutivism. The core idea here is that some normative feature is used to justify some constitutive feature(s) of agency, and then properties of these normatively justifiable features are used to explain other normative phenomena. Hence, the kind of agency that explains *some* norms is normatively justified from the start, but can still do explanatory work regarding *other* norms. More generally:

[12] Beyond the three views just mentioned, the first hint of such a strategy is arguably to be found in Bagnoli (2013, p. 11), though she does not develop the point in detail.

Moreover, Paakkunainen (2018) has recently suggested an interesting view in the vicinity of those I discuss here. Her response to the shmagency objection is to say that reasons can be grounded in features of agency that we need not instantiate. But it does not seem committed to taking those features to be normatively justifiable or inescapable for us, so it differs slightly from the new solution to the problem that I will present – my solution is formulated in terms of a kind of normative inescapability.

Nevertheless, her positive suggestion for how one might formulate a form of constitutivism that avoids the shmagency objection relies on a kind of attributivism about goodness, making it very similar to Smith's view. I suspect that *any* version of constitutivism that relies on saying that we need not descriptively instantiate the constitutive features of agency properties of which explain normativity will have to involve some normatively inescapable feature, and hence will be susceptible to the problem I will raise for it.

> (*PARTIAL CONSTITUTIVISM*) A form of constitutivism is a form of partial constitutivism iff the constitutive features of some aspect of agency properties of which explain normative phenomena are normatively justifiable (or desirable, required, etc.) to instantiate, rather than only descriptively necessary to instantiate for one to instantiate that aspect of agency.

I call partial constitutivism 'partial' because, on this view, it is not the case that all norms are explained by constitutive features that an agent only descriptively instantiates. Instead, at least one norm is a deeper feature of the explanation, suggesting how some aspect of agency (or just 'agency', for short) is justified. That type of agency becomes normatively, not descriptively, inescapable. This norm (or these norms) may or may not be further reducible to descriptive constitutivist or naturalistic terms – but whether it (or they) can be is an open question that one need not take a stand on.[13]

Partial constitutivism stands in contrast with standard forms of constitutivism in at least two ways. First, it is in one sense normative rather than descriptive. The properties constitutive of some feature(s) of agency by which norms are supposed to be normative are normatively justifiable (or desirable, required, etc.), and that is why one should instantiate them. They may or may not also be descriptively necessary for agency – one may have to live up to the standards of agency in some minimal sense to count as an agent at all. But then, those normative standards in turn impose stronger norms on an agent, e.g. to be a fully functional agent (cf. Smith's view). Or, alternatively, there might be some sort of external normative justification for being an agent of the relevant kind. This kind of justification does not require agency to be descriptively inescapable at all (cf. Bratman's and Arruda's views).

By contrast, according to standard formulations of constitutivism, such as Korsgaard's, Velleman's, and Katsafanas', the norms constitutivism explains are explained by non-normative properties that constitute something as a member of their kind *simpliciter*, where the kind is 'agency'. One must instantiate them to at least some extent to count as an agent, but the constitutive features themselves need not put direct normative pressure on agents. Whether one should be an agent is a separate question from whether one is one.

Second, partial constitutivism is less comprehensive than many standard forms of constitutivism. It does not attempt to explain all practical norms or all moral norms, but

[13] However, at least Smith (2017) believes that agency, *qua* goodness-fixing kind, is so reducible.

rather uses some normative feature to explain agency, which in turn explains some other normative phenomena, such as moral reasons. This might seem to make it less ambitious, and therefore less attractive, than the standard forms of constitutivism. But the view still does substantive work to explain some normative phenomena (e.g. moral reasons) in terms of others (e.g. perfectly good agency), and so remains informative.

As mentioned, partial constitutivism can be exemplified with Smith's view, as presented in chapter 2, section 5. To repeat: on his present view, our reasons for action have their sources in the desires of our ideal counterparts, where our ideal counterparts are perfectly good *qua* agents. These counterparts serve as our ideal advisors. Here, agency is taken to be a goodness-fixing kind, and a perfect exemplar of an agent, Smith thinks, is fully rational, and rationality is spelled out in terms of coherence.

To explain reasons, then, Smith appeals to a prior conception of goodness for an agent. The goodness here is functional; a perfectly coherent agent is a perfectly functioning agent. Furthermore, Smith thinks that functional goodness should be understood in terms of features that are constitutive of agency. Again, as mentioned, it does not matter for theoretical purposes whether the functional goodness here is constitutivist as long as one can get to the explanation of reasons in terms of the responses of perfectly functioning agents. That is still what Smith gives us: one can explain reasons in terms of the desires of perfectly functioning, or perfectly rationally coherent, agents, where that type of functional perfection is understood independently of our reasons.

So Smith is a partial constitutivist. He takes fully functioning agency to be able to explain normative reasons, and it is good for us to be such agents because agency is a goodness-fixing kind. Doing so, like other partial constitutivists, he is able to deny premise (1) rather than (2) in the original shmagency argument. For according to partial constitutivists, it is irrelevant whether or not actual agents are ideal. Their reasons can be explained regardless, for the kind of agency that explains norms is not descriptively inescapable, but normatively inescapable – on Smith's view, this is because it is good to be a perfectly functioning agent. Importantly, it does not even matter whether the reasons that one attempts to explain are the reasons of a member of some kind of entity that does not instantiate agency. Even the reasons of Martians or Saturnians can quite possibly be explained by the responses of idealized agents.

Hence, on a partial constitutivist view, it is possible to explain moral norms or other forceful reasons without requiring the agency in terms of which they are explained to be descriptively instantiated. This is because it is *idealized* agency which explains the norms, not actual agency. On this view, questions about whether we should be agents or shmagents are first-order normative questions about which reasons we have, but the reasons themselves have a deeper explanation (cf. Smith, 2015). So partial constitutivists seem able to explain at least some normative phenomena without invoking the argument against premise (2) discussed in the previous sections.

(6) Shmagency as Underdetermination

Unfortunately, partial constitutivism lends itself to another version of the shmagency objection. In sections 3 and 4 above, I defended premise (2) in the shmagency argument. The external shmagency question still stands if constitutivists cannot explain shmagents' reasons. But I just argued that partial constitutivists can deny premise (1) instead, for we need not be ideal agents descriptively, only normatively, and ideal agency remains the same regardless of who we are, descriptively.

However, a problem for (2) remains even when it is re-interpreted by appealing to normatively inescapable agency. This is a problem of *underdetermination*. The reasons or other norms that we can get out of some form of justifiable (or desirable, required, etc.) agency do not seem normatively preferable to other, slightly different, potential reasons or norms that we can arrive at by treating the same justified (or desirable, required, etc.) agency slightly differently. There seems to be little reason to prefer one theory of norms explained by normatively justified or inescapable agency to another, for the reasons that are to be captured in such terms can be given different interpretations depending on the extent to, or manner in, which we should treat such agency. And there are many possible ways of baking the same normatively justified form of agency into our norms, immediately yielding many possible candidate theories about them.

This problem exists because it is unclear how we should treat the normatively justified form of agency that explains norms for us. For, very plausibly, there are different ways to respond to the normative feature(s) that it might have, which gives rise to different possible sets of norms that are captured by different forms of agency. For example, if the normative feature is some form of goodness or value, we can ask: ought the goodness or value of agency be maximized, satisficed, used to explain only some subset of our reasons (or other norms), respected, honoured, promoted, or something else? Until we have an answer to this question, we can always ask: 'Why should I care about norms explained by ideally rather than shmideally justified agency?', where the difference between an ideal and a shmideal agent is that we treat their goodness or value differently.

We can exemplify this problem, too, using Smith's view. Why should we be concerned with any of the alleged reasons that the desires of our ideal advisor counterparts supposedly grant us in virtue of functioning perfectly? There is, on his view, some sense in which it would be good to be a fully functioning agent, and still good – but possibly less good – to be slightly less perfect as an agent. But no reason is provided for thinking that we should care more about reasons that have their sources in the desires of a perfectly functioning agent than a slightly less perfectly functioning agent. Being the latter type of agent may not be as attributively good for us as the former, but their responses may still be what gives us reasons. So until we know what to do with the value of agency, it seems like it is not just ideal agents, but also less than ideal agents, who have desires that are decent candidates for explaining our reasons. This gives rise to the shmagency question: why should we care about what an ideal, rather than a shmideal, advisor desires?

This type of problematic underdetermination appears in several places in Smith's framework. One form stems from the possibility of a satisficing conception of how we should handle the value of agency. It is possible that someone who seeks advice would be happy with advice from an agent who is good enough rather than one who is ideal. We are free to attribute functional perfection to all kinds of things without for that reason thinking that we are normatively required to care about having the best versions – if a perfectly sharp knife is an ideal knife, I do not need an ideal knife to cut myself a piece of bread, just one that does the job. Similarly, a merely good enough agent might have desires that are good enough to count as sources of reasons.

Second, there does not seem to be any reason to pick out Smith's ideal advisors rather than several other advisors even if we want perfect advice. A less than fully coherent, yet still idealized, version of me might have desires for (*A*) lasagna, (*B*) mac and cheese, and (*C*) spaghetti carbonara, but desire lasagna over mac and cheese (*A* > *B*), mac and cheese over spaghetti carbonara (*B* > *C*), and spaghetti carbonara over lasagna (*C* > *A*). These desires are intransitive, and therefore incoherent. But that version of me might also always desire spaghetti Bolognese more strongly any of the other dishes, leaving the intransitive desires moot, because – assuming, in this case, that my reasons vary with the strength of my desires – I will always have more reason to cook spaghetti Bolognese rather than any of the other dishes. It is not obvious that we ought to prefer Smith's view to this one when it comes to accounting for the reasons we have.

There are potential replies here. Smith (2017) thinks that the desires of ideal agents grant us reasons because they are authoritative, unlike other desires. They are preferable to the desires of less than ideal agents for two reasons. First, he thinks, we tend to find out what we have reason to do by deliberating. Second, the extension of what we have reason to do is well captured by what we would be motivated to do if we deliberated well.

It is unclear how these points have bearing on the underdetermination worry, however. We may well find out what we have reason to do by deliberating well *enough*, and the extension of what we have reason to do may well be fixed by what we are motivated to do if we reason well *enough*. So the satisficing worry stands. Similarly, intransitivities among desires we do not actively consider need not trouble our deliberation in practice, and hence the extension of what we have reason to do may well be compatible with cases such as the pasta case. It is very plausible that an ideal version of me would not even consider making lasagna.

Another potential reply is that we might want to do more normative thinking prior to going constitutivist. Then we can, maybe, settle how we ought to treat the valuable agency in terms of which we explain reasons, and then go on to defend some form of constitutivism based on the kind of agency we believe to be valuable and know how to treat.

I admit that this strategy seems open, but the reply essentially concedes the problem. We need to explain how to handle the goodness of agency to avoid the underdetermination worry. There is more work to do to explain what we are supposed to do with the value of agency before we can explain things in terms of it. And partial constitutivism is an incomplete view until we know how to treat valuable agency, and therefore which norms it is supposed to explain.

(7) Conclusion

What have we learnt from the preceding discussion? In section 1, I presented the shmagency objection, and in section 2, I showed that the standard line of defence – i.e. that the shmagent is self-defeating – seems to hold firm. But then I argued, in sections 3 and 4, that the shmagency objection remains because there still is an external standpoint from which there are reasons (or something like them) that constitutivists cannot handle. After that, in section 5, I argued that partial constitutivists can defend themselves against the argument by denying premise (1) rather than premise (2). But, in section 6, I argued that this response fails due to a shmagency objection which stems from underdetermination.

From this discussion, one might conclude that constitutivism still fails due to the shmagency objection. But that would also be the end of this dissertation, and it is not. Instead, we may use the preceding discussion to develop *desiderata* that a viable form of constitutivism must make sense of. A viable form of constitutivism must not succumb to the new shmagency worries I have presented.

There are, more specifically, two key *desiderata* here. First, (*i*), if one is after explaining reasons, one must be able to explain reasons for the kind of sophisticated shmagents I discussed in sections 3 and 4. Second, (*ii*), one must be able to avoid the problem of underdetermination. I shall show how my preferred form of constitutivism handles these *desiderata* in chapter 7.

4. A Humean Theory of Agency

The last two chapters have mostly been destructive. In chapter 2, I argued against the leading forms of moral constitutivism in the literature – they suffer from versions of the problems of substantive assumptions and adequacy. And in chapter 3, I argued that they suffer from two new shmagency worries.

But my aim is to formulate a form of constitutivism that can avoid these problems. More specifically, I aim to formulate a version of constitutivism based on a theory of agency built from elements that stem from the Humean theory of motivation (*HTM*), viz. the theory that (at least) a belief and desire, suitably linked up, make events actions. My idea is to defend a theory of agency based on *HTM* because of the intimate connection between action and agency. If we can understand action, we can then develop a theory of agency out of the constituents of our theory of action – and that theory can later be put to use to explain various norms. Hence, my strategy will be to argue from action to agency, and, later, from agency to norms.

How can I do that? *HTM* has, famously, received severe criticism over the past decades, so it will not do just to assume it.[1] Instead, in this chapter (and in the next, and in appendix A), I intend to defend it. I will argue that *HTM* plausibly is true for a at least one central class of intentional actions. This is sufficient to make it very plausible to think that agents are constituted by the properties that constitute events as intentional actions according to *HTM*, whether or not some actions require more or less than that.

Still, this defence of a Humean conception of agency will most likely strike some as inconclusive. Dissertations could be, and have been, written on *HTM* alone (e.g. Petersson, 2000; cf. Heuer, 2001). But I shall focus on its constitutivist pay-offs, and for that, it suffices to tip the dialectical balance in my favour. I shall attempt to do so by, first, providing a positive argument for the view, and by, second, replying to objections to *HTM* in chapter 5 and appendix A. This makes *HTM* – and its associated conception of agency – tentatively acceptable.

The game plan, then, is to start in section 1 by explicating my preferred way of understanding *HTM*. In section 2, I present what I call the argument from centrality, according to which central cases of action ought to be explained by a version of *HTM*. In section 3, I extend that argument to show how it makes us justified in accepting a

[1] This is Wallace's (2003) line of response to Joyce (2001)'s *HTM* assumption.

Humean conception of agency. I conclude in section 4, showing how I think the Humean conception of agency improves on the constitutivist views I criticized in chapter 2.

(1) The Humean Theory of Motivation

> When I 'raise my arm', my arm rises. And now a problem emerges: 'What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?' (Wittgenstein, 2009, p. 169)

Wittgenstein's question is often used to introduce a central question in contemporary philosophy of action: what distinguishes raising one's arm as an action from raising one's arm as a mere reflex or bodily movement? Or, in other words, how do we distinguish an action, with some (usually, but not necessarily, physical) extension, from an event with the same extension? The question can be generalized to cases of action that do not involve movement. We may ask: what is the difference between acting and merely behaving?

It is actions of the kind that differs from mere behaviour that I shall be concerned with here. But, even if we draw that distinction, the question still needs further refinement. It is possible that there are actions that are more akin to 'mere' behaviour than what we are aiming to bring out with the distinction between acting and such behaviour. It might be the case that, for some actions, there is nothing of interest left when one has subtracted the fact that one's arm goes up from the fact that one raises one's arm.

To avoid such worries, I shall focus on explanations of *paradigmatic action*.[2] A paradigmatic action (for humans) is one that is both a statistically common instance of an action (for humans), and one which sets a normative standard for action. Hence, I shall assume that paradigmatic actions are *successful tokens* of actions that are typical everyday actions for humans.[3] A few (somewhat egocentric, but still typical) examples of such actions include grabbing a beer from the fridge because it will quench my thirst, going out for some fresh air just because I feel like it, buying a painting for my living room because it looks good, and writing this dissertation in the couch at home instead of in my uncomfortable office armchair since that is more comfortable.

[2] Dancy (2018) calls a version of this theoretical approach, which starts out with fundamental cases as the most central ones, 'focalism'. I extend my use of this method significantly in chapter 5, section 7, and appendix A.

[3] Strictly speaking, however, this point does not matter yet – but I shall treat the issue of unsuccessful actions at length in chapter 5, section 7.

The reason I assume that paradigmatic actions are successful is that, while they may be normatively bad in many ways, if there are norms or aims inherent in actions, only actions that are successful with respect to the norms or aims inherent in those very actions can be paradigmatic actions. And this, in turn, is because unsuccessful actions fail to live up to at least some aspect of the norms inherent in the very actions they are – unsuccessful actions, as I understand them here, are just the actions that fail to live up to those norms or aims. This means that they are not even successful versions of the kind of actions they are themselves, and hence hardly can set normative standards for actions.

Furthermore, the cases of action I have in mind are individual (rather than collective), involve taking means to ends, and are intentional.[4] Perhaps we need extra assumptions – e.g. intentions, individual or collective, construed as distinct mental states – to explain sophisticated actions, but I conjecture that complex (individual) actions will also involve the features of paradigmatic actions.[5]

Similarly, it might be true that less complex machinery is needed to explain some less sophisticated actions. Perhaps there are intentional actions which are less central cases of action than paradigmatic actions, and which should not be explained in the same way.[6] Or maybe there are non-intentional actions, or even intentional behaviour that is not sophisticated enough to count as action. I shall presume that it is unproblematic that a Humean agent can do these things along with standard, intentional actions – for example, if a Humean agent has beliefs and desires so that she can perform means-ends actions, and there are actions out of desires that do not involve means-beliefs, she can plausibly perform such actions just because she has desires. Or if actions out of emotions are not reducible to actions out of belief/desire-pairs, perhaps emotions should be added to her psychology. That is a very minor tweak of the view.

Can there be other paradigmatic actions than individual, intentional, means-ends actions? Perhaps even some of the other actions I just mentioned are paradigmatic too? There are two ways to reply to this worry. A hard-line answer is to hold that it, indeed, is the case that all paradigmatic individual human actions involve taking means to ends and are intentional. One could point at some property of means-taking and argue that, whatever else paradigmatic action features, it must feature at least that.

[4] However, I make no commitment to what explains their being intentional. They need not, for example, be intentional because they are performed for reasons.

[5] So why have I not I started by talking about such complex actions instead? Well, they seem to either be based on too strong assumptions (cf. chapter 2) or we can easily avoid acting in such ways (cf. chapter 3).

[6] Typical examples of such actions are actions performed out of emotion (Hursthouse, 1991), skilled actions (Ruben, 2003, ch. 4), and actions performed out of habit (MacIntyre, 2016, ch. 1).

There are many options for what that property might be. One possibility is intelligibility (cf. Smith, 2012*b*). It is possible that means-ends structures are necessary for actions to be intelligible to the agent (or others). Another possibility is that there might be some level of control that means-ends actions involve but other actions need not. Perhaps the agent is awkwardly heteronomous if she does not, herself, take the means to some end. A third possibility is that the agent may be too passive to count as acting if she does not contribute to the action by doing something for some purpose. A fourth possibility is that the strategic thinking involved in means-ends actions might be central to a distinctively human life form. So there are different ways to spell out hard-line replies.

A soft-line answer, on the other hand, is to allow for distinctions inside the concept of the 'paradigmatic.' We can allow for central and non-central types of paradigmatic actions. If so, perhaps for the reasons just mentioned, it is plausible that means-ends actions are a more central case of paradigmatic actions than other actions, even though these other actions also may count as paradigmatic. If so, central paradigmaticality will plausibly matter more than paradigmaticality *simpliciter* for action-theoretical purposes, and one may then reformulate my argument in terms of central paradigmaticality. Nothing turns on whether one prefers this term, however.

I shall presume that some reply, either hard-line or soft-line, is successful. With that point in mind, I shall now proceed to provide a theory about at least one type of (individual) intentional means-ends action. As per my Canberra-influenced methodology, a theory of such actions must capture our pre-theoretical intuitions about how such actions work. To that end, I shall argue that paradigmatic actions work according to *HTM* – a belief and a desire, suitably linked-up, non-deviantly causing an event are a necessary component of what makes an event such an action.

More specifically, I shall defend:

> (*CENTRAL HUMEANISM*) A necessary feature of what makes one central class of events intentional actions is that a belief/desire-pair, suitably linked up, non-deviantly is part of the cause of the events in question.

A number of comments are needed. First, as mentioned, I will limit my discussion of what makes actions 'actions' to talking about one central – or paradigmatic – class of actions. Some actions may need less than being non-deviantly caused by suitably linked up belief/desire-pairs by way of explanation, and some actions may need more. But regardless, I take beliefs and desires to be fundamental features of what makes this central class of actions 'actions'. This means that *CENTRAL HUMEANISM* is a more minimal claim than most formulations of *HTM* (e.g. Davidson, 1980*b*; Hempel, 1961; Sinhababu, 2017, ch. 1; Smith, 1994, ch. 4; 2009; 2012*b*). With suitable extra assumptions, it may however be developed into those views.

Second, one can probably fine-tune the ontology of *CENTRAL HUMEANISM*. I have written that belief/desire-pairs *make* events actions, so actions are actions in virtue of their causes. Strictly speaking, I would like to go with an even stronger Humean view: my preferred interpretation of *HTM* is to *identify* (at least one central class of actions) with events caused non-deviantly by (at least) belief/desire-pairs. Then we can understand this kind of actions as a generalization over token instances of events caused in that way. But I use the making-language because it is weaker; it is possible that being an action (of the kind under discussion) is a distinct property from being that kind of event, so talking in this way makes fewer assumptions.

Moreover, I do not even want to wholeheartedly commit to events-talk. While most forms of *HTM* tend to be events-based, *CENTRAL HUMEANISM* might possibly be formulated in terms of some other underlying ontology, such as a process-based one or even an agent-causal one. Events-talk is simple and conservative, and therefore what I prefer, but I trust the reader to make adequate substitutions here if she wants to.

Third, one may also want to fine-tune the mental states that feature in *CENTRAL HUMEANISM*. Perhaps the terminology of beliefs and desires ought to be changed; maybe knowledge rather than belief is the most fundamental factive mental state (Williamson, 2001, ch. 1), and hence a feature of action-explanations. Or maybe intuitions, construed as seeming-states, can be part of action explanations. Or maybe 'pro-attitudes' is a better term than desires. And it might even be misleading to talk about belief/desire *pairs* as opposed to belief/desire *sets*, for it is possible that instrumental desires are composed out of more than one single belief/desire-pair. Again, I have stuck with standard formulations for simplicity and conservativeness, but I trust the reader to make adequate substitutions.

Unfortunately, in spite of these clarifications, it is still unclear exactly what *HTM* contains and what kind of explanation it provides. I shall conclude this section by saying something more about that. Remember: the core idea of *HTM* is that at least part of what makes an event into an action is that it is non-deviantly caused by a belief and a desire, suitably linked up. It is therefore – on my view – a species of the causal theory of action; events are made into actions in virtue of their causes.

Moreover, I will follow most Humeans in taking beliefs and desires to be functionalistic mental states, to be understood in part in terms of their ability to cause action when linked up with each other. Desires explicate what one wants to do by means of representing a proposition one aims at ('having an aim'), and beliefs what one takes the world to be like, *inter alia* in order to reach one's aims ('having a means'), but also epistemically because they plausibly aim at truth, knowledge, or justification. The set of states that can count as desires here is, however, very broad – it features *all* motivational states. And when a motivational state is suitably linked up with a belief in a belief/desire-pair – or, in other words, in an instrumental desire – it may then cause an action.

I have also added an anti-deviant causation-clause to *CENTRAL HUMEANISM*. This standard move is there to rule out the infamous problem of deviant causation, viz. the problem of explaining what happens when we cause events if something interferes with our causing them, but that thing then causes them. We can bracket this problem for now (but cf. appendix A, section 2, for discussion).

While I now have introduced the components of *HTM*, one may still wonder what kind of explanation *HTM* might yield. There are versions of *HTM* that are causal, and versions that are not. On non-causal interpretations, the belief/desire-pairs that make events actions play the role of making the actions intelligible, where 'intelligibility' means something like how the action makes sense to the agent, e.g. because it is an appropriate outcome given what the agent takes into account before acting (cf. O'Brien, 2018). The belief/desire-pairs are then supposed to make actions intelligible by *rationalizing* them as motivating reasons for action – but without causing the actions. I shall, however, stick with the standard causal view. While it does not rule out rationalizing explanations – as we shall see, it supports them – its core commitment is that belief/desire-pairs causing events in the right way are at least part of what makes them actions.[7]

Here, David-Hillel Ruben (2003, pp. 90; 192-193) attempts to distinguish between causal theories of *action* and causal theories of *action explanation*. Explanations, he thinks, are epistemological, whereas causal theories of actions – such as *HTM* – are metaphysical; they concern the grounding relations between facts. But this distinction seems unfair. What causal theorists of action have – or at least ought to have – in mind is that when a

[7] A potent argument for the causal view is a dilemma. Either action explanations are causal, or they are not. If they are, non-causalism fails. But if not, non-causalists face 'Davidson's challenge' of explaining how actions can be brought about without causation (Mele, 2003, pp. 38-45; cf. Davidson, 1980*b*).

certain kind of mental cause of an event occurs, e.g. if a belief/desire-pair causes an event in the right way, the event is made into an action *because* it is caused by a belief/desire-pair. If that is some sort of metaphysical (or 'ontic') explanation, a case of grounding that causal theorists of action sometimes just have called an explanation, or something else seems to be splitting hairs.[8] What matters is that, using the terminology from chapter 1, section 2, according to which *F* explains *G* if *G* is so because *F* is so, what happens here is that an event *G* is made into an action *because* there is a causal relation *F* between a suitably linked up belief/desire-pair and *G* (and the causal relation is non-deviant).

Even so, one may wonder how the *causation* of action looks here, especially as *HTM* is in part teleological due to the functionalist interpretation of the mental states it features. The idea is that a desire sets an aim that an agent then brings about via her belief. But despite being in that sense teleological, *HTM* can be combined with many different theories of causation. It might be that the causal explanations straightforwardly involve belief/desire-pairs in hypothetico-deductive arguments (Hempel, 1961). Or it might be that if a desire has the functional role of motivating the agent to achieve some end, this aim may still fundamentally be explained by some more complicated causal organisation working on a lower level of explanation (e.g. in the brain) (Davidson, 1980*a*). Or it may involve a counterfactual folk concept of causation (Ruben, 2003, ch. 6). Or something else. And, importantly, *contra* 20th-century writers like Hempel and Davidson, it need not be the case that causal laws are involved; generalizations are fine. Here, too, the reader is free go with her preferred interpretation.

Finally, in addition to causing and making events into actions, belief/desire-pairs *also* do one more kind of work. At least in the ordinary case, they rationalize action. An action is explicable as intelligible to the agent (and others) because it stems from her beliefs and desires. As such, belief/desire-pairs are ordinarily interpreted as 'motivating reasons' for the action – an action is motivated by the belief/desire-pair, and it is because of that reason that agents act (though cf. appendix A, section 3, for complications).

But now to the argument.

(2) The Argument from Centrality

The argument for a Humean conception agency appeals to the centrality of means-ends actions in human life. It has two conclusions. In this section, I shall defend *CENTRAL*

[8] For Skow (2016), explanations are answers to 'why?'-questions. That seems true here as well.

HUMEANISM, and hence the first conclusion of the argument. In the next, I extend it with an extra premise about agency, yielding the second conclusion.

The argument goes like this:

> (1) Paradigmatic actions involve agents taking means to ends, in the manner of *HTM*.
>
> (2) If paradigmatic actions involve agents taking means to ends, in the manner of *HTM*, then paradigmatic actions are actions in virtue of being events caused by (at least) a belief/desire-pair, suitably linked up, that cause the events in the right way.
>
> ---
>
> (C1) Paradigmatic actions are actions in virtue of being events caused by (at least) belief/desire-pairs, suitably linked up, that cause the events in the right way.
>
> ---
>
> (4) If paradigmatic actions are actions in virtue of being events caused by (at least) belief/desire-pairs, suitably linked up, that cause the events in the right way, then one is epistemically justified in taking beliefs, desires, and their links to be partially constitutive of agency.
>
> ---
>
> (C2) One is epistemically justified in taking beliefs, desires, and their links to be partially constitutive of agency.

I take premise (1) to be fairly easy to establish. Already in my characterization of paradigmatic actions, I appealed to several cases that plausibly should be cashed out as means-ends actions. These were: grabbing a beer from the fridge because it will quench my thirst, going out for some fresh air just because I feel like it, buying a painting for my living room because it looks good, and writing this dissertation in the couch at home instead of in my uncomfortable office armchair since that is more comfortable.

It is easy to see how these actions involve taking means to ends. I grab a beer *because* I am thirsty, go for some fresh air *because* I feel like it, buy a painting *because* it looks good, and decide to write in my comfortable couch at home instead of in my uncomfortable office armchair at work *because* I want to be comfortable because I desire pleasure. Here, I desire to quench my thirst, do what I feel like, to own a good-looking painting, and pleasure. Combined with relevant means-beliefs, the actions I perform to satisfy these desires can explained by belief/desire-pairs.

In fact, Humean actions like these seem to be ubiquitous, both in folk- and more scientific psychology. One kind of evidence for thinking of desire-based actions as common is that it is hard to deny the influence of desires on many types of action. Many desires have an obvious phenomenology that make people take means to satisfying them – Sinhababu (2009) emphasizes 'the cases of hunger, thirst, and sexual lust' (p. 485) as especially clear examples of desires that few, not even non-Humeans, deny cause actions (together with beliefs).[9] Such desires have a motivating force that is phenomenologically obvious, so they seem part of what explains actions based on them – and more often than not they explain actions when combined with beliefs. The beer-drinking case is telling here. Thirst gives me a desire to drink, I know there is a beer in the fridge, and hence I go there to grab it.

The point can be extended beyond 'primal' desires like the three just mentioned. What philosophers call 'desires' need not just be what the folk call desires. The class of relevant mental states also includes whatever is meant with the attitudes mentioned in phrases such as 'I feel like φ-ing' or 'I want to φ' or 'I have an urge to φ' – they indicate that we have something motivational going on inside us, which pushes us to act. Whether or not the folk would think of these kinds of motivation in terms of desire, the everyday cases that fall under those headings should also be incorporated under the heading of 'desire' in the Humean motivational functionalist sense of the word. A good example is that I may go for some fresh air to clear my mind just because I feel like it, and I know that going outside for some fresh air is a way of doing so.

There are also various kinds of more scientific cases that are hard to make sense of without appealing to desires. This is not to say that there are no competing interpretations to the desire-based interpretations of the cases, but insofar as they interpretations are at all plausible, the argument for explaining at least some paradigmatic actions in terms of desires is strengthened. I shall provide some interpretations here that show how my view unifies a greater number of cases than the folk-psychological ones.

For a theoretical case, much research in the social sciences is incomprehensible without reference to preferences. Preferences, in turn, should at least in some cases be interpreted in terms of how much we desire the outcomes. There are two main theories of preferences (Thoma, ms.). One is broadly realist. It interprets preferences as mental states. These unproblematically qualify as desires. They are pro-attitudes, and that is just what Humeans ordinarily have in mind with desires.

The other theory is revealed preference theory, according to which preferences are read off from people's choice behaviour. Here, whatever one chooses in some choice

[9] He especially mentions that they do so because of how they are produced and how little reason has impact on them, but it is also true that most tend to experience them.

situation is what one prefers (of the alternatives in the choice situation). It follows that preferences are outcomes of choices, so whatever machinery explains one's choice explains one's preference. But even on such views, desires plausibly explain preferences in at least *some* cases. This is because it seems very hard to explain how one could choose something if that choice is not based on something motivational in the first place. It makes little sense to think that I would prefer to spend money on a painting because it looks good instead of a flowerpot which I think looks worse unless I desire to decorate my living room with something that I think looks good.

There are also more empirical examples of cases where desires seem to do work. To take just one, in social and cognitive psychology, there is a recent and influential literature on goal frames – ways of understanding and acting in a situation based on a deeper conception of the situation and a motive, where the motive sets a goal which frames what one does (Lindenberg and Steg, 2007; cf. further references in the paper). For example, if 'having pleasant experiences' is one of my motivating goals, it could drive me to write from home rather than in the office. And that motive may very obviously be interpreted as a desire.[10] If we accept that there are at least some such goal frames, we should accept that there are desires underlying them. These desires tend to be linked up with beliefs about how to satisfy them – and then we satisfy them.

The different kinds of means-ends actions I have discussed here amply support premise (1). Hence, there is strong reason to think that we should interpret paradigmatic actions in terms of their being motivated by belief/desire-pairs. Are there objections? Yes, of course. But I defend *HTM* against objections in chapter 5 and appendix A, so I refer the reader there for them.

Hence, premise (1) seems acceptable. Premise (2) is *prima facie* trickier, however, because I will not try to argue for a complete theory of action based on *HTM*. It is just a necessary condition of paradigmatic actions. Nevertheless, there is good reason to think that many actions are actions in virtue of having been caused in the right way. This is because it seems well-established in both folk- and scientific psychology that many actions can be explained using the standard belief/desire pattern of *HTM* – that is what premise (1) established. And if that is the case, one may wonder: why would anything else (that does not include belief/desire-pairs) be what explains these actions?

[10] Another alternative would be to interpret it as a preference, but given the arguments above, that would plausibly just turn it into a type of desire.

This point can be formulated as a challenge to any alternative theory that excludes the features of *HTM*.[11] If so, given how these causal action explanations go, *HTM*, in the form of *CENTRAL HUMEANISM*, seems like a reasonable dialectical starting point in any theory about what (standard, human) action is. It does not seem like more machinery than that which features in these causal action explanations for actions to take place – and if we can explain why an action takes place using (at least) Humean belief/desire-pairs, then it seems very plausible that what makes something an action should include at least the causing of an event by (at least) a belief/desire-pair.

I have not established that *CENTRAL HUMEANISM* can explain all actions, however. It might be thought that this leaves its explanatory power, or indeed other theoretical virtues, fairly limited. A more comprehensive theory seems like it would be better. If so, how can I remain confident in *CENTRAL HUMEANISM*?

If pressed, I would not really mind accepting a more comprehensive form of *HTM*. It does not seem implausible to me that all (individual, intentional) actions might be explained using (at least) the machinery of *HTM*, but that view is far too controversial to be defended in the limited space I have available here. Fortunately, however, my move is free-standing from that commitment. It seems overwhelmingly plausible that some actions should be explained using *HTM* for the reason just given, even if not all actions are. At least paradigmatic actions should be explained by it, so *HTM* is plausibly part of any fuller theory of what actions are.

Another response to my argument is to claim that while I may have established that *HTM* is the dialectical starting point – hardly surprising; it is also known as 'the standard story of action' – it does not follow that *HTM* is correct. To establish that, I would need to defend *HTM* against objections. Again, however, I try to do so in chapter 5 and appendix A. So for now, I regard (C1), inferred from (1) and (2), as *prima facie* justified.

(3) From Action to Agency

To go from (C1) to (C2), and hence the Humean conception of agency, we need to establish premise (4). Premise (4) goes from the existence of paradigmatic Humean actions to our being epistemically justified in believing in a Humean theory of *agency*. Here, I say 'epistemically justified in believing' rather than just inferring the Humean theory of

[11] By contrast, it is not a challenge to a theory that includes *HTM* but also something else.

agency from the conception of action because I do not think the argument in favour of the Humean theory of agency *necessitates* the theory; rather, it ('just') makes it very plausible.

How does it do that? *CENTRAL HUMEANISM* can support a first version of a theory of agency that is enough for developing a form of constitutivism. This is because a move similar to the one that supports premise (2) can be made again, but about agency. As beliefs and desires are part of what makes Humean actions 'actions', one may ask: need there be anything else, or at least anything else that excludes beliefs and desires, that makes agents 'agents'? Hence, *HTM*-based actions generate a challenge to all conceptions of agency that are not based on (at least) beliefs and desires. Absent a convincing reply to it, plausibly, a theory of human agency should include at least these Humean features.

Moreover, a Humean theory of agency can also explain many other things that we take agency to involve – and do so well. First, it is a platitude that an agent (of some kind) must have all the properties involved in the initiation of actions (of the kinds that they have the properties to be able to initiate) when they initiate actions, in the sense that these properties make what they do an action. This is because the agent is whatever stands behind an action. And this we can explain by saying that a Humean-style agent must have at least beliefs and desires.

Similarly, agents are usually, in some sense, larger than any particular actions they perform. Agents usually extend over time, and usually are able to perform more actions than a single one. (This is a key feature of Korsgaard's argument against particularistic willing, cf. chapter 2, section 2.) This is easy to explain on a Humean view. The agent can consist of, at least, all the beliefs and desires that a person has at a particular moment in time, and plausibly also over time. Any version of a theory of agency should allow for this kind of diversity.

Furthermore, one can be split in one's agency. For example, one can feel torn about what to do, act on desires one does not endorse, or regret not satisfying some desire while satisfying other, incompatible, ones. (This point, too, is emphasized by Korsgaard in her defence of the argument against particularistic willing in chapter 2, section 2.). There is nothing strange about this either on a Humean view. One person's agency can involve many disparate beliefs and desires (cf. Sinhababu, 2017, ch. 10).

Accordingly, a Humean theory of agency can capture all these pre-theoretical points while still looking reductive, in the sense that it does not appeal to any extravagant features or properties. It can make sense of the pre-theoretical phenomenon of agency regardless. Hence, we should be justified in accepting it.

But how, exactly, does it look? Because it is based on the Humean explanation of at least one central class of paradigmatic actions, I shall let the Humean theory of agency be called *PARADIGMATIC AGENCY*. I therefore propose that a necessary, and indeed constitutive, feature of the kind of agent who can perform paradigmatic actions is:

> (*PARADIGMATIC AGENCY*) A person *P* is a paradigmatic agent, i.e. the kind of agent who is able to perform paradigmatic actions, only if *P* has a relevant set of beliefs and desires, and *P* is such an agent at least partially in virtue of having them.

An agent, then, is a person who is at least in part constituted as an agent in virtue of having beliefs and desires. I shall also assume that, in virtue of having these mental states, an agent has the capacity to hold those states fully functionally, and to act on them. Since an ideal agent is constituted by her psychology, that capacity is also constitutive of her.

Note also that *PARADIGMATIC AGENCY* will be expanded in chapter 5 – for now, just be aware that it is a starting point, and many amendments may have to be made to it. For example, I assume that all agents are persons, but leave the notion of a 'person' unanalyzed here. Hence, the reader should feel free to introduce her favourite theory of personal identity as an underlying factor in the view, or even discard talking about persons if she denies that assumption or is a nihilist about personal identity. As usual, I trust her to make adequate theoretical substitutions.

Moreover, perhaps one should add further conditions to *PARADIGMATIC AGENCY*. It is plausible that a person counts as an agent only if the beliefs and desires she has are such that they can be combined to cause actions, so they cannot be too disparate or incoherent, or that her beliefs are means-beliefs, in line with some Davidsonian notion of how rationality may be a constitutive ideal of interpretation of the mental (Davidson, 1980*a*). Furthermore, maybe the agent's second-order desires must be aligned with her first-order ones (Frankfurt, 1971), or maybe she must be able to make value judgements, construed in terms of certain kinds of desires (Street, 2008; Taylor, 1985) or beliefs (Smith, 1994, ch. 5). In the light of all these possibilities, I have formulated *PARADIGMATIC AGENCY* in terms of a 'relevant' set of beliefs and desires. But regardless of what is the case here, however, *PARADIGMATIC AGENCY* provides an outline of a theory of agency, and that is enough for the argument from centrality to be completed.

Or is it? I did allow that there might be non-intentional 'actions' (back in section 1), and *CENTRAL HUMEANISM* is compatible with actions demanding even less than the features involved in the explanation of paradigmatic, intentional actions. For example, there might be actions out of, or from, emotions. If nothing more is needed to explain a type of agency than the fact that agency is constituted by the properties needed to explain some kind of actions, then what is so bad about saying that agency is constituted by the properties needed to explain *such* actions?

Two things. First, I am interested in paradigmatic agency. This means that non-paradigmatic actions, or at least non-central cases of paradigmatic actions, are outside the scope of what I am after. Second, one may admittedly think that there may be agents who do not act from belief/desire-pairs, but are only able to act in other ways, such as out of emotions. But if we were to allow the existence agents without at least beliefs – or other mental states that aim to represent the world as it is – alongside motivational states, there would be agents without any beliefs whatsoever. And it seems extremely implausible to think that one could be an interesting kind of agent without any beliefs whatsoever. Then one could not, even in principle, take means to one's ends, but any interesting kind of agent should be able to do that. So the argument from centrality is safe for now.

(4) Conclusion

I have defended *PARADIGMATIC AGENCY* by appealing to the argument from centrality, which itself is based on a defence of *CENTRAL HUMEANISM*. Moreover, the Humean theory of agency avoids the problems that beset other constitutivist explanations of action and agency. It avoids the problem of substantive assumptions by having been given an independent defence – which I, naturally, take to be adequate. Therefore, it also avoids the problem of adequacy. And, importantly, it does so while simultaneously being able to avoid the problems for constitutivism that I presented in the previous chapter. I will conclude this chapter by showing how it does that.

First, the problem I raised for Korsgaard's view is that she assumes that agents have to be fully unified, in the sense that they have maxims universalized by *CI*, to act. Clearly, Humeans are not committed to that. Being able to explain the possible disunity of agency is one of the arguments that speak in favour of *PARADIGMATIC AGENCY*.

Nor have I appealed to maxims or universalization, which makes this view preferable to Korsgaard's due to its weaker action-theoretical commitments. In all these respects, *PARADIGMATIC AGENCY* is better than Korsgaard's view.

Second, when I introduced Velleman's view (in chapter 2, section 3), I argued that interpretations of it that rely on desires or drives for self-knowledge seem implausibly strong. There is no need to appeal to such desires or drives, and indeed, *PARADIGMATIC AGENCY* does not. However, the main problem I presented for Velleman's view is a problem for his take on intentions; he runs control of actions and self-knowledge together too tightly. Here, Humeans can construe intentions as at least either belief/desire-pairs, or add more, such as Bratmanian functional states (Bratman, 1999), Anscombean dispositions to be able to answer 'why?'-questions about what one is doing (cf. Anscombe, 1957), or something else entirely. Regardless, no directive beliefs that run control and self-knowledge together are necessary.

Third, the problem for Katsafanas' view is that he is committed to saying that all actions aim at *POWER*, and hence to explaining them in terms of drives. Being based on a Humean explanation of action, *PARADIGMATIC AGENCY* does not suffer from this shortcoming. No actions need be guided by drives at all on this picture – though some may if drives are just uninteresting dispositions to cause desires. Hence, the problem for Katsafanas can be avoided.

Fourth, finally, the problem for Smith's view is that it is implausible to think that agents must have dominant desires to help and not interfere to avoid wishful thinking. The view I have defended is clearly not committed to such desires. Of course, a Humean view may be formulated so that it will require some desires – in chapter 7, I shall argue that ideal agency in fact is partially constituted by a desire to cooperate with other cooperative agents. But I am not committed to Smith's argument.

Hence, *PARADIGMATIC AGENCY* improves on the leading constitutivists' views. It avoids, I think, the problems of substantive assumptions and adequacy, for it is based on weak yet independently defended assumptions. In the coming chapters, I will try to show how it can be developed into a two-tiered form of constitutivism.

5. Instrumental Rationality and Its Normativity

By now, I have argued against standard constitutivist explanations of action and agency (chapter 2; 3) and instead started to defend a Humean one (chapter 4; appendix A). A theory of agency based on *HTM* supersedes older constitutivist views.

I have also indicated that I aim to develop a two-tiered form of constitutivism where the first tier involves a defence of a principle of instrumental rationality and its normativity, and the second tier a defence of a take on practical reasons. In this chapter, I present the main positive argument for the first tier.

To this end, I shall first argue that belief-desire explanations of action have to be supplemented with a principle of instrumental rationality. (I will use the terms 'instrumental principle', 'principle of instrumental rationality', 'norm of instrumental rationality', 'requirement (of instrumental rationality)', and sometimes the label (*IR*), interchangeably below to name it.) With this principle, we can explain how agents stand behind their actions. Second, I shall also argue that this is not just any principle of instrumental rationality, but one that has normative force. Using this extension of the Humean theory of agency, I shall then proceed to develop the second tier of my Humean constitutivism in chapters 6 and 7.

Here, however, section 1 sets the stage for my argument. I explicate my strategy to show what theories of rationality and normativity must do for the purpose of my project, and also situate my aims in the contemporary literature on rationality. In section 2, I explicate the notion of instrumental rationality that I work with.

In the key section 3, I then re-develop Korsgaard's argument against particularistic willing (cf. chapter 2, section 2) into an argument suggesting that instrumental rationality is constitutive of agency. While she demands that agents precede their actions with a procedure by which they universalize their maxims, a similar but weaker procedure, where agents link beliefs and desires together by the instrumental principle, can solve the problems her account has but still explain how the agent may stand behind actions. In section 4, I also show how this account improves on more established solutions to the problem of the disappearing agent.

In section 5, then, I develop a constitutivist-friendly conception of normativity. In section 6, I show how the principle of instrumental rationality is normative in that way. I then reply to the problem of bad action in section 7, and summarize the chapter in section 8.

(1) The Problem of Normativity

Constitutivism is supposed to explain why we are subject to norms as well as why they have normative force. And what I fundamentally am after in this dissertation is to explain the normativity of moral norms – in particular, moral reasons. Whatever else moral or other practical reasons are, they are normative or prescriptive; they have so-called 'prescriptive force' or 'normative oomph' (Joyce, 2001, ch. 2). However, because I aim to explain reasons in terms of the desires of idealized agents, to explain the normativity of reasons, I need to show how reasons based on such desires are normative. There is nothing about desires, by themselves, that make them seem apt to be the sources of reasons, so my practical reasons risks ending up normatively arbitrary unless I say more about them (cf. e.g. Heuer, 2004; Korsgaard, 1996*a*, ch. 3).

Here, I shall do some groundwork that will allow me to say more by trying to explain a kind of normativity that goes deeper than the normativity of reasons. More specifically, I shall argue that the normativity of rationality should be understood in terms of *directivity*. Rationality directs our actions, regardless of what we aim to do. And as I will be arguing that instrumental rationality is prescriptive (because it is directive), I shall be able to argue that reasons explained by the desires of (*inter alia*) instrumentally rational ideal agents are *legitimized* (cf. chapter 6, section 2). So I shall be able to explain how reasons are prescriptive.[1]

I shall, however, start with the normativity of rationality. To put the explanation in terms of my chapter 1 constitutivist schema, I shall argue that the normative phenomenon instrumental rationality (P_P) has its normativity explained in terms of properties (C) constitutive of a feature of an aspect of paradigmatic agency (A^*) – namely, paradigmatic agency itself.

However, the literature on the normativity of rationality has exploded since John Broome (1999) started writing about rational requirements. Even an incomplete list of key texts would have to feature at least (Broome, 2013; Fink, 2014; Henning, 2018;

[1] The reader may have noticed that my framework looks a lot like James Dreier's Humeanism about reasons (1997), so I want to clarify how it relates to my strategy. His argument, roughly, is that instrumental rationality is necessary for action, which is close to how I think, but also normative bedrock. Only if agents are instrumentally rational can they, in virtue of being instrumentally rational, have reasons. There is a clear similarity between our views, for rationality lies at the heart of both his view and mine.

But the views are not the same. Instrumental rationality, on my construal, is one of the properties of an *idealized* agent whose hypothetical desires explain the reasons we have – I shall even argue that my view plausibly can explain reasons for shmagents who are not rational (cf. chapter 7, section 4). Dreier, on the other hand, makes no reference to idealization when it comes to reasons. For him, agents have reasons because they are instrumentally rational from the start.

Hussain, ms.; Kiesewetter, 2017; Kolodny, 2005; 2007*a*; Korsgaard, 2009; Lord, 2018; Levy, 2018; Parfit, 2011*a*; Raz, 2005*a*; 2005*b*; Schroeder, 2014; Smith, 2009; 2012*c*; Scanlon, 2007; Southwood, 2008; 2018*b*; Wedgwood, 2017). Comprehensively handling this vast literature is impossible here. But I shall attempt to locate the proposal I have in mind in this literature, as well as motivate its position, here and in appendix B.

To start off, then, there are some different kinds of theories about the nature and normativity of rationality in this literature. The received view is that there is a set of principles of rationality, and that the principles (or principle, if there just is one) are normative (e.g. Fink, 2014; Hussain, ms.; Korsgaard, 2009; Lord, 2018; Schroeder, 2014; Southwood, 2008; 2018*b*; Wedgwood, 2017). This set need not contain the same principles according to all writers, but most think that it includes at least some version of a principle of instrumental rationality, some enkratic principle suggesting that one ought not to be weak-willed, and some principles of theoretical rationality such as one suggesting that one should not believe both p and $\neg p$ at once. There are exceptions, however; for example, Korsgaard believes that *CI* explains instrumental rationality, where that is understood in terms of the hypothetical imperative (2009, pp. 68-70).

Other philosophers agree that there are principles of rationality, but do not think that they are normative (Kolodny, 2005; 2007*a*; Raz, 2005*a*; 2005*b*). They are so-called 'myth theorists'. They believe that while there are principles of rationality, they have no interesting normative force. Instead, all their seeming normative work is done by reasons, so, for example, the seeming normative role of instrumental rationality is captured by instrumental reasons. Interestingly, Broome himself hesitates to take a stand on the question of whether requirements (i.e. principles) of rationality are normative. He writes that he would like to think that the principles of rationality are normative, but that he has yet to find a convincing argument for this thesis (Broome, 2013, ch. 11).[2]

More recently, some other philosophers have denied that there are structural principles of rationality in *any* sense (e.g. Henning, 2018; Kiesewetter, 2017). However, they tend to agree with the myth theorists that all the seeming normativity of rationality can be fundamentally explained by the normativity of reasons. Hence, on their views, too, rationality tends to consist of responding to normative reasons – and the reasons do the relevant normative work, though there are no distinct requirements or principles of rationality at all.

[2] Some other philosophers even hold mixed views, where rationality is normative in some very weak sense but something else has a deeper kind of normativity (cf. Parfit, 2011*a*, ch. 5; Ridge, 2014, ch. 8).

However, it still seems pre-theoretically plausible to think that there is a set of prescriptive principles of rationality. First off, we often talk about (in particular) instrumental rationality as a principle. This indicates that thinking of it in that way is a starting consideration – and that is why it is the received view. Moreover, the received view is also still that rationality has normative force in such a way that it differs from arbitrary desires, feelings, or wants. It applies to all paradigmatic actions, and seems more constraining than desires, feelings, or wants do. Moreover, because rationality plausibly comes in the form of principles, its force is *not* plausibly the same as the force of reasons. Reasons aggregate with other reasons, principles do not.

Because rationality is normatively forceful, moreover, it seems like we can be criticized if we fail to be rational (cf. e.g. Parfit, 2011*a*, ch. 5; Kiesewetter, 2017, ch. 2; Ridge, 2014, ch. 8). But, notably, criticizability does not come before normative force in the order of explanation; we are plausibly criticizable when we are irrational *because* rationality is normative. And this kind of criticism seems to come in addition to the kind we might receive if we fail to act for our reasons – if an agent gets all her reasons wrong but still fails to take the means to what she, falsely, thinks her reasons recommend her to do, she seems criticizable for that, quite independently of what she has reason to do (cf. Kauppinen, forthcoming). Furthermore, because rationality is normative, it should both be possible to fail to live up to it (ought-implies-can-fail; cf. Korsgaard, 1997) and to live up to it (ought-implies-can; cf. Southwood, 2018*b*).[3]

I aim to defend a principle of instrumental rationality which captures all the properties I have suggested that it seems to have in the last two paragraphs, and I will present a constitutivist argument for this view. There is already a burgeoning literature on constitutivism about rationality (cf. Bratman, 2016; Broome, 2013, pp. 204-205; 2008; Coons and Faraci, 2010; Fink, 2014; Goldman, 2011; Kauppinen, forthcoming; Kolodny,

[3] Michael Ridge (2014, ch. 8) provides another list of platitudes that he thinks a theory of rationality must capture. These are: (*i*) rational agents can set and abandon ends, (*ii*) rational agents must have wills to set an end, (*iii*) to have wills, rational agents must be able to ask themselves what they are to do, (*iv*) it is up to rational agents what they are to pursue, (*v*) rational agents with impulses must be able to reject or delay satisfying them, (*vi*) rational agents can take the essential means to their ends because they believe the means are essential, (*vii*) rational agents can will whatever they take to be the means to their ends, (*viii*) rational agents can revise their ends when they take those to conflict, (*ix*) rational agents can form new ends to specify more abstract ends, and (*x*) rational agents are capable of making and acting on normative judgements.

There is no tension between this list and the account of rationality I shall develop below; indeed, my account of rationality is very similar to Ridge's, though our views on its normativity diverge.

2005; Korsgaard, 1997; 2009; Levy, 2018; Mylonaki, 2018; Railton, 1997; Southwood, 2008; 2018*b*; Smith, 2009; 2012*c*), but this argument is novel.[4]

However, I shall leave other possible requirements of rationality, structural or not, aside. Hence, the instrumental principle need not be the full story about practical rationality – let alone about rationality in general. Nevertheless, my limited claim about instrumental rationality still gives us a very explanatorily powerful view (cf. chapter 6; 7).

This matters, for the view has a lot of work to do. In general, it needs to (*a*) handle the pre-theoretical commitments we seem to have to rationality and its normativity that I mentioned above. For my purposes, it also needs to (*b*) be able to partake in an explanation of the normativity of reasons. Moreover, the conclusion of chapter 3 featured two *desiderata* it needs to make good on to feature in a plausible constitutivist framework. *Desideratum* (*i*) – that a theory of reasons should explain reasons for shmagents – is not very important here, for I am not yet discussing reasons. However, *desideratum* (*ii*) – that one must avoid a problem of underdetermination – needs a solution. I shall show how my form of constitutivism about instrumental rationality can do all this.

(2) Instrumental Rationality

However, to defend a theory of instrumental rationality, I need to say something about how I understand it. As was the case with *HTM*, there is no doubt a dissertation to be written on instrumental rationality itself, so many questions about its nature will have to go unanswered here. The reader is free to answer most of these – apart from those I will cover, for I will have to answer some questions.

First, then, generally speaking, a principle of means-ends rationality implies that there is a form of irrational incoherence involved in an agent's failing to combine her ends and means. Hence, it does *not* need to say anything about maximizing preference satisfaction or prudence (cf. Korsgaard, 1997). An extended principle can do so, but I have a weaker notion of rationality in mind.

[4] Note, however, that Smith (2012*b*) briefly suggests a defence against the problem of the disappearing agent which is similar to mine (cf. section 3; 4 below). Replying to Velleman's (1992) version of the problem of the disappearing agent (cf. chapter 2, section 2, footnote 5), Smith holds that an agent's capacity for instrumental rationality is what helps her connect the dots between her ends and intentions, and between intentions and actions.

However, this solution is not an argument for the principle: Smith's defence of his view rests on his own controversial (2009) defence of a capacity for instrumental rationality as a solution to the problem of deviant causal chains (cf. Mayr, 2011, pp. 117-121). By contrast, I argue that we should think that *HTM* should be supplemented with a principle of instrumental rationality *because* this helps it solve the problem of the disappearing agent.

We can use the following generic formulation to capture the intuition behind the means-ends principle:

> (*IR*) If *A*'s end is to φ, and *A* believes that ψ is a necessary means to φ, and *A* does not take the means ψ, then *A* is instrumentally irrational.[5]

With this negative formulation *IR* in the background, we can then go on to say that *A* is instrumentally rational only if she is not instrumentally irrational (but not iff, because arationality is possible).

Unsurprisingly, however, just how *IR* should be formulated is controversial. For now, the reader should feel free to use her own favoured formulation, as long as it involves saying that there is a kind of irrationality involved in not being responsive to one's means-beliefs given that one has some end. Since I work with Humean background assumptions, I will use this formulation:

> (IR_{HUMEAN}) If *A* desires to φ, and *A* believes that ψ is the best means to φ, and *A* does not have an instrumental desire to ψ, then *A* is instrumentally irrational.

IR_{HUMEAN} is a first Humean stab at capturing the kind of coherence that I shall argue that an agent must have to act. The core idea is that it requires the agent to have combinations of desires (generating ends) and beliefs (about means) that she unifies into instrumental desires, which then may cause actions.

Some comments might be helpful here, even though IR_{HUMEAN} is a first stab. First, it is formulated in terms of beliefs and desires. It is based on the Humean theory of agency defended in the last chapter, where at least one class of paradigmatic actions are explained, minimally, by at least a belief and a desire. It then explicates the principle of instrumental rationality in terms of how such a principle might look given these psychological assumptions, hence fleshing out the general theoretical framework about action and agency that I defend here.

Second, the principle is formulated in terms of 'desires', 'beliefs', and 'best means.' These formulations will probably have to be altered before we arrive at a final statement

[5] This version of *IR* is adapted from Kiesewetter (2017, p. 15). The most important difference is that I have replaced his 'intending' with 'ends' to generalize it even further – an intention is an end, but there can be other ends as well.

of the principle. Forming instrumental desires based on all one's desires does not seem rationally required – normatively poor or just very weak desires need not count. Nor will I say anything about what happens when an agent's beliefs are false.

Instead, for now, just assume that agents have some sets of desires and true means-ends beliefs, and that some of these desires are the agents' *ends* (cf. Hubin, 2001). This means that they are the desires that are relevant to satisfy for the agents, on some interpretation of 'relevant' – again, normatively poor or just very weak desires need not count. Accordingly, a fully formulated principle will probably not require that the agent forms belief/desire-pairs based on all her desires. Hence, when I refer to desires that an agent is required to satisfy, I mean to refer to desires that are ends. Sometimes, I shall even call them 'end-desires' for the sake of clarity.

Moreover, the 'best means' formulation of IR_{HUMEAN} also from Kiesewetter's formulation in terms of necessary means. Not much turns on this terminology, however; any account of rationally necessary means will have to be extended to explain what should happen when several possible means are available, and any theory formulated in terms of 'best means' will have to say what that means to be complete. In fact, formulations of *IR* in terms of 'necessary' and 'best' means are in the same boat in cases where there only is one necessary means to take – then it is also the best one.

But what about cases where there are many means available? In such cases, I will just assume that the 'best means' is, rationally speaking, 'the means to take' in a way which is analogous with 'the thing to do' (cf. Gibbard, 2003; Ridge, 2014; Southwood, 2018*b*). This means that I take it to be the rationally best means to take, whatever 'best' is, for the agent in her circumstances. It is possible to plug in any theory about what the best means is here, e.g. whether it is the most effective one, a satisficing one, one that is fitting for the sake of reaching the agents' ends, or something more complex. What view one goes with here should not matter for broader theoretical purposes.

Third, recent debates about the nature and normativity of rationality have often focused on a number of logical properties. Most importantly, it is unclear whether requirements of rationality should be understood as wide-scope requirements – in the case of *IR*, governing entire means/ends-pairs, allowing one to drop the end if the means is unacceptable – or narrow-scope requirements – requiring one to take the means given that one has an end (Broome, 2007; Kolodny, 2005; Schroeder, 2014). But there are also other issues in these debates, such as whether rationality governs states or processes

(Kiesewetter, 2017, ch. 3; Kolodny, 2007*b*), or whether it is synchronic or diachronic (Ferrero, 2012; Hedden, 2015; Lenman, 2009).

Though IR_{HUMEAN} may have to be reformulated to capture whichever of these properties one prefers, I shall largely ignore these debates for now. It is unclear how all these properties should be understood in the first place (cf. Kiesewetter, 2017, ch. 3). And, more importantly, it turns out that the wide-scope/narrow-scope debate is irrelevant for constitutivist purposes. The result that rationality is normative will apply regardless.

That conclusion follows from the following argument (cf. Katsafanas, 2013, pp. 50-53). Rational requirements are either wide-scope or narrow-scope. If they are narrow-scope, whatever they require will follow on a means-ends take on rationality, for it just specifies that one needs to combine the best believed means with one's ends on pain of irrationality. Given IR_{HUMEAN}, if one desires to φ, believes that ψ is the best means, and is not arational, then one is rationally required to link φ and ψ together.

But if rational requirements are wide-scope, it might seem like one can escape taking the means to an end by changing one's end (or desire). However, *any* paradigmatic action a Humean agent can perform will have an end set by another desire, for paradigmatic actions involve Humean-style desires. If so, the instrumental principle will remain binding, for one cannot perform any paradigmatic action without it.[6] Hence, it does not matter whether I appeal to wide-scope or narrow-scope norms. The normativity of instrumental rationality is not undermined because one can have different ends.

Hence, I will push discussions of the logical properties of rationality aside and proceed with the main issues. To recapitulate after this lengthy introduction, these are: establishing that instrumental rationality is constitutive of agency (section 3; 4), that it is normative because it is constitutive of agency (section 5; 6), and extending the Humean understanding of action in the light of the problem of bad action (section 7).

(3) Developing Korsgaard on Unity

The reader may remember Korsgaard's argument against particularistic willing from chapter 2, section 2. In my critical discussion, I focused especially on her premise (3), viz. the conditional saying that 'if the agent is a unified whole, then she wills according to universalized maxims.' However, even though premise (3) is flawed, there is still

[6] What about non-paradigmatic actions? I shall argue, in section 6 below, that paradigmatic actions are plight inescapable, so non-paradigmatic actions do not matter for now.

something to Korsgaard's argument. In the light of *PARADIGMATIC AGENCY* from chapter 4 and IR_{HUMEAN}, I shall redevelop it for my own purposes.

Here is my reformulation:

(1) If an agent is to perform a paradigmatic action, then she precedes her action with a procedure by which she makes herself into a cause in the world.

(2) If an agent precedes her action with a procedure by which she makes herself into a cause in the world, then she initiates her action by going through a procedure in which she links up beliefs and desires via the instrumental principle (which also makes the action into a paradigmatic action).

(3) If an agent is to perform a paradigmatic action, she initiates her action by going through a procedure in which she links up beliefs and desires via the instrumental principle (which also makes the action into a paradigmatic action).

(4) If an agent initiates her action by going through a procedure in which she links up beliefs and desires via the instrumental principle (which also makes the action into a paradigmatic action), then the instrumental principle is part of the best constitutive explanation of paradigmatic agency.

(C) If an agent is to perform a paradigmatic action, the instrumental principle is part of the best constitutive explanation of paradigmatic agency.

Premise (1) is almost completely Korsgaard's. Both her worry about the disappearing agent and her core line of response seem fundamentally correct to me. There are some differences between her way of approaching to the topic and mine, however. A first minor difference is that I have written 'paradigmatic action', using my notion of paradigmatic action from chapter 4, sections 1 and 2. This stands in contrast with Korsgaard's view, as she limits her argument to human rather than animal action instead.

A much more important point is that we interpret the problem of the disappearing agent slightly differently. In chapter 2, section 2, I wrote that Korsgaard thinks the problem has two aspects. The first aspect is descriptive; we must explain how agents can stand behind actions and actively bring them about. The second aspect is normative; agents can succeed or fail in ways that ordinary causes cannot, and we must be able to respond appropriately or inappropriately to actions because agents are responsible for them.

However, I think the problem of the disappearing agent only is descriptive. The reasons she offers for thinking that the problem necessarily has normative import are weak. She first mentions that a certain class of success verbs (e.g. 'succeed/fail') apply to agents, but not other phenomena, indicating their normative standing. I am not sure about whether that is true, but even if it is, I do not see why it would have any metaphysical implications. Our language can hardly show that agency is normatively *constituted*, as opposed to merely being such that norms apply to or have force for agents.

More importantly, her point about attributing actions to agents for the sake of being able to hold them responsible seems both normatively question-begging and misdirected. First, as different first-order ethical theories have very different implications about normative ethics, we should not assume much, if anything, about how responsibility works prior to knowing at least something about which first-order ethical theory we should go with. But just what kind of theory we can establish is what constitutivism is supposed to show (cf. chapter 1, section 2).

Moreover, even if we were to take our practices of responsibility attribution at face value, the attribution of responsibility to a diachronic *agent* rather than a diachronic *person* seems off the mark. The question of whether someone deserves punishment for some past crime if she has changed in significant ways since her crime seems to concern whether she still is the same *person*, not whether she is the same *agent*. But agency and personhood should not be identified (absent an argument for that). Admittedly, with *PARADIGMATIC AGENCY*, I have assumed that agents are persons, but it is possible to hold any mainstream theory of personal identity – e.g. bodily views, psychological views, further facts views (Parfit, 1984, pt. III), or even narrative views (Velleman, 2006) – without thinking that those theories say much about agency, and vice versa. So it is plausibly personhood (on some relevant interpretation), not agency, that needs to be unified for the purpose of explaining our practices of responsibility.

With these points in mind, I believe the best interpretation of the problem of the disappearing agent is the descriptive one. But even if we stick with that interpretation of it, there are at least two *desiderata* that a theory of agency has to meet: the agent has to contribute to her action actively, and she must stand behind her actions.

We can appropriate Korsgaard's strategy to explain these *desiderata*. The best constitutivist explanation of them involves saying that there is something active that the agent must do to generate an action; this is how she makes herself into a cause by going through a certain procedure. The agent *stands behind* the action by contributing with a causal pattern prior to its performance, and she is *active* because she must *do* that prior to acting (though the doing here is not an action itself).

This solution to the problem of the disappearing agent is rather theoretically virtuous. It is not intrinsically tied up with Korsgaard's view; it is in principle open regarding which mental states it involves, and it is even compatible with a very neat, event-causal, conception of agency – including *PARADIGMATIC AGENCY*.

On to premise (2). Here I start to deviate much more substantially from Korsgaard. Her second premise involves a notion of unification that is absent from my premise (2). This is because unification does not play as significant a theoretical role for me as it does for her. Instead, I will explicate my notion of the pre-action procedure by appealing to the formation of instrumental desires via a linking up-process between beliefs and desires using the instrumental principle. (Of course, one may call forming an instrumental desire a kind of unification, but that should not be confused with Korsgaard's stronger version of unification.)

The picture where the principle is necessary for linking up beliefs and desires has two steps. First, linking beliefs and desires is needed for paradigmatic actions, i.e. cases where the agent succeeds in taking the means to their ends.[7] Second, because beliefs and desires are different mental states, there has to be a third factor linking them together to generate instrumental desires. Whatever brings them together is the third factor and is simultaneously what the agent contributes to her action. And a principle of instrumental rationality (e.g. *IR_{HUMEAN}*) either is, or can be understood as a representation of, that third factor. In the latter case, it is strictly speaking not the *principle* that is the third factor, but rather the agent's capacity to follow it. However, since the capacity has to be isomorphic to the principle, this view does not differ in any interesting way from the former one. I shall proceed to speak of these views interchangeably below.

To defend this view, then, I will proceed by first presenting the picture I have in mind more clearly, and then give an argument for it. I will also show how it does better than some alternative Humean psychologies, and finally go on to show how it solves the problems I have presented for Korsgaard's premise (3). It is here that the Humean solution improves on her view.

The general picture is this. Beliefs and desires are different kinds of mental states. Moreover, having both of them need not entail that they automatically become an instrumental desire consisting of a belief/desire-pair, for they do not put themselves together automatically. If I know that I can go to Berlin by flying there, and I suddenly

[7] Note that it is from here and on that it matters that I treat paradigmatic actions as successful instances of actions (cf. chapter 4, section 1, esp. footnote 3).

acquire a desire to go to Berlin, I do not necessarily connect these mental states just because I form the desire. Something has to happen for them to align; they must link up to turn into an instrumental desire (cf. Schueler, 2009; Smith, 2009).

But then, some sort of linking-up process seems needed for us to be able to stake out courses by which we may act. It is here that agents can manifest a form of activity by creating a pattern in the causal order, and hence actively bring an event about while also standing behind it. This is because agents can manifest their capacities to bring means and ends together by forming an instrumental desire out of a belief and a desire. That manifestation is the right kind of activity to be the contribution of the agent, and as the belief/desire-pair the agent has generated – in the instrumental desire – then is part of the causal etiology of some action, the agent will stand behind the action (by having manifested her capacity to form the instrumental desire that caused it) after having contributed actively to it (by bringing the belief/desire-pair together). Hence, this view solves the problem of the disappearing agent.

We get the following three stages in the etiology of paradigmatic actions. The stages describe some different possible configurations of how the different states in the agent's psychology can fit together, and hence generate an action (if correctly set up):

> (*STEP 1*) The agent has unattached beliefs and desires. They are not linked up. However, she also has a principle (or capacity) for instrumental rationality. Insofar as her beliefs and desires are not linked up, she is unable to perform a successful paradigmatic action, just because these states are not linked up into a pair.
>
> (*STEP 2*) The belief/desire-pair gets linked up by the principle of instrumental rationality. If successful (e.g. non-akratic), the agent now has an instrumental desire, generated by the manifestation of her principle (or capacity) for instrumentally rationality. This means that the agent is in a position to act.
>
> (*STEP 3*) The linked-up belief/desire-pair will, absent further complications (e.g. *akrasia*), cause events that count as actions in virtue of having the linked-up pair as a cause.

There is more to say about how the linking process works. First, it need *not* be a case of reasoning (or deliberation, assuming these are the same), choice, or intention-formation, though it can stem from them. Schueler (2009) states, without argument, that what goes on in Humean linking up-processes usually is a kind of reasoning, and moreover seems to hold that it must be phenomenologically occurrent. This is supposed to lead Humeans to a dilemma: either the linking up-process is occurrent or not. If it is, then Humeans, implausibly, seem to take practical deliberation to standardly involve basing one's deliberations on one's desires. If it is not, then there is no process.

But there is no reason to think that the linking process must be or involve a reasoning process as long as Humean agents are rationally sensitive to the conclusions of deliberation and construct belief/desire-pairs in line with whatever that is. What matters is just that the agent is able to link beliefs and desires up – for example, in accordance with her reasoning. So the actual linking need not involve a process of reasoning or say anything about the agent's phenomenology.

Moreover, because the operative rational factor here amounts to a kind of rational sensitivity, *mutatis mutandis*, the linking process need not say anything about choice or intention-formation. Forming an instrumental desire is no doubt closely related to that, but one may form belief/desire-pairs without choosing to act on them, and I am silent on the relation between intention and belief/desire-pairs.

Furthermore, not any old form of linking will do here. One can think hypothetically or theoretically about what to do without for that reason come to actually form instrumental desires. I can represent some set of beliefs and desires, whether or not they are mine, and think about how they can be put together, without for that reason doing anything with my actual desires.

So what happens when the agent links beliefs and desires up? Regardless of how we understand the ontology of following the principle, a Humean agent may have rational capacities (cf. Smith, 2003*a*), i.e. coherence-inducing patterns of responses to possible configurations of her psychology. Here, the relevant one involves being more or less means-ends coherent over different possible worlds because different counterfactuals are true of the agent depending on whether or not her beliefs and desires can be linked up as specified by the principle, or not. Then the agent can put her mental states together in virtue of manifesting her capacity for instrumental rationality (or fail to do so).

The linkage provided by instrumental rationality can be either synchronic or diachronic. Instrumental desires generated by the linking process can be satisfied at any given moment when they are formed, or may take a much longer time to satisfy, depending on the content of the beliefs and desires that are involved. We can act on some occasion without any commitment to acting in a similar way in the future if we have desires that can be instantaneously satisfied and beliefs about how to do so immediately. So instrumental rationality can be synchronic. However, many desires also take time to satisfy, and the beliefs involved in taking the means to satisfy them can be conditional on future events, too. Hence, instrumental rationality may also be diachronic. In such cases, one keeps acting on the instrumental desire one has over time, or it starts making one act once the belief it features has become true.

So much, then, for the description of what goes on. By why do we need this kind of link? That one needs to link up beliefs and desires to act paradigmatically is already implicit in the considerations just adduced. Absent some link between beliefs and desires, we cannot act on any particular desire that we have, because it (at least usually) remains causally idle unless it is combined with a belief. That is the point of the Berlin case. Such linkage does, however, not by itself require instrumental rationality.

But there is a special way in which belief/desire-pairs must be linked up to cause successful actions – namely, by the instrumental principle. The agent can link a belief/desire-pair up into an instrumental desire, which by itself forms part of the etiology of an action when it causes some event. With the principle, this linkage is due to something that the agent contributes, i.e. a manifestation of it that is part of a procedure which generates a certain causal pattern. So rather than just being the effect of external phenomena, a coherence-inducing principle can combine beliefs and desires.

Manifesting a capacity to follow a principle like that seems like a prime Humean candidate for how agents actively can contribute to bringing actions about while simultaneously standing behind them. Because agents play at least these roles, and Humeans can explain how this may be the case in these ways, we have reason to think that Humean agents are, in this sense, instrumentally rational.

By now, I have characterized and defended a conception of the instrumental principle. But one might think that there still are some Humean psychologies according to which the principle seems unnecessary. I will exemplify this problem by criticizing three other Humean views. Their failures indicate – though they do not conclusively show – that adding a principle remains necessary for the right kind of linking process, for it seems like there is no other plausible candidate for a mechanism that can do it.

First, there is the view that beliefs and desires can be interdefined as being such that they cause action independently of third factors. Beliefs and desires are functional states, and among their functions are to cause actions together with the other kind of mental state. One might think that they are individually necessary and jointly sufficient as – or at least two constituents of a sufficient explanation of – the right cause that constitutes events as actions. Here, there is no need for an extra principle of instrumental rationality. This is a very common view among Humeans (cf. e.g. Sinhababu, 2017; ch. 2; 5; Smith, 1994, ch. 4).

But I do not doubt the general functionalist picture. Among the functions of beliefs and desires, respectively, is that they can initiate actions when linked up with the other. But nothing follows about *how* they have to be linked up from this characterization. The suggestion here is that to be active so that she can stand behind her actions, an agent links up her mental states up via a third factor. So we should characterize the functions of desires and beliefs as being such that they link up with the other mental state to produce actions via IR_{HUMEAN}.

An alternative Humean view is Davidson's (1980*a*; 1980*b*; 1980*c*; cf. Smith, 2009, for discussion). Davidson does not think that a principle of instrumental rationality features in the belief-desire pairs that constitute motivating reasons. Rather, instrumental rationality should be understood as a background constitutive principle of agents' psychologies that can be read off from belief-desire combinations that we attribute to agents, so the combinations are prior to the principle in our attributions of it to the agent. This is an aspect of Davidson's broader anomalous monist programme. He thinks there is a constitutive ideal of rationality, including instrumental rationality, that we must attribute to agents to be able to explain their mental lives, because without it, we cannot explain errors (Davidson, 1980*a*). But if so, what matters for action explanations may just be the belief/desire-pairs from which we can read the explanation of particular action tokens, not whether the principle does causal work to link up any given pair.

Davidson's concerns are orthogonal to mine, however. The basic point that beliefs and desires do not link up automatically stands unchallenged if one assumes the primacy of belief-desire combinations absent a principle uniting them. But more than that is needed for an action; for the agent to actively bring it about, she needs to precede her action with a certain procedure. But Davidson offers no explanations of how *that* may work; rather, his theory posits a background principle, not one that needs to be active in any particular case. Adding an element of activity to agents would complete his view.

Third, one might think that beliefs and desires can be linked up because desires direct attention (Sinhababu, 2017, ch. 5). The core idea here is that one of the things desires do is to direct one's attention to their outcome, so means-beliefs about how to satisfy them become motivationally relevant insofar as one has one's attention directed at how to satisfy one's desires.

But here, too, it is unclear why and how beliefs and desires issue in actions in the right way unless there is some third factor linking them up. It is not enough to say that desires direct attention and therefore join up with beliefs, for there is nothing interestingly active the agent does when she has her attention directed to a belief by a desire. To make a Korsgaardian point, here, the mental states become causal factors *within* the agent, rather than something the agent can use to act. On pain of the agent disappearing, this view needs to be supplemented with a story about how actions are issued based on something the agent does rather than having some of her mental states causing actions – and plausibly, what an agent does is to apply a principle of instrumental rationality.

My premise (2), then, looks defensible. I can now show how it solves the problems I presented for Korsgaard's argument against particularistic willing in chapter 2, section 2. The first is Millgram's (2011) worry that it seems implausible that it would be necessary to form maxims that are universalized in Korsgaard's sense, requiring one to stick to them in all possible situations that are relevantly similar, just to maintain a distinction between the agent and one particular action. One might, for example, have policies that last over many situations without needing them to be universal instead. However, belief/desire-pairs can be completely particular, though they can often also involve desires that take plenty of time to satisfy.

The second problem for Korsgaard is that we might not need Kantian maxims, let alone universalized ones, at all in our descriptions of action to explain them. This suggests that her view involves unnecessary theoretical machinery. But here, there is no need to form universalized maxims, for agents need not form maxims at all. Nor is there any need to have universalized belief-desire combinations either, because there is no need to universalize one's maxims (or belief-desire combinations).

Premise (2) is defended, then. (3) follows straightforwardly. Moreover, I have also already hinted at a defence of premise (4) in my defence of *PARADIGMATIC AGENCY* in chapter 4. I argued that a good way to explain several features of paradigmatic agency is to posit that it is explained by the properties needed to initiate action. In particular, we can explain how people are agents reductively by appealing to their beliefs and desires. This yields an explanation of agency in terms of:

> (*PARADIGMATIC AGENCY*) A person *P* is a paradigmatic agent, i.e. the kind of agent who is able to perform paradigmatic actions, only if *P* has a relevant set of beliefs and desires, and *P* is such an agent at least partially in virtue of having them.

But now I have argued that instrumental rationality, as well, is necessary to account for paradigmatic action. We should then extend the Humean conception of agency to include instrumental rationality in the explanation of paradigmatic actions, so the instrumental principle features along with beliefs and desires in the explanation of paradigmatic agency. (Similarly, like above, I shall also assume that an agent with principle of instrumental rationality has a (constitutive) capacity for instrumental rationality – perhaps the capacity *is* the principle, or at least she has a capacity to act based on it.) We get:

> (*PARADIGMATIC AGENCY$_{IR}$*) A person *P* is a paradigmatic agent, i.e. the kind of agent who is able to perform paradigmatic actions, only if *P* has a relevant set of beliefs and desires as well as *IR$_{HUMEAN}$*, and *P* is such an agent at least partially in virtue of having them.

With (4) established by the defence of *PARADIGMATIC AGENCY$_{IR}$*, (C) follows straightforwardly. So there is an argument for a linking-up process based on instrumental rationality which stems from Korsgaard's original argument for *CI* based on unification. And this argument can avoid the problems that beset her original argument.

(4) The Reappearing Agent

Here is the picture: paradigmatic actions stem from instrumental desires. These consist of belief/desire-pairs, linked up by the instrumental principle. Because beliefs, desires, and instrumental rationality are constitutive of paradigmatic agency, a paradigmatic agent stands behind the way the beliefs and desires are linked up, and she links them up actively by exercising her instrumental rationality. So there is an agent behind paradigmatic action. This means that the causal chain leading up to action has to involve a certain procedure: a linking-up procedure.

The explanation of the latter point lies in how the agent creates a causal pattern to generate it. Because the properties that constitute paradigmatic action also explain paradigmatic agency, it is the properties of the agent that, when made manifest, explain an instrumental desire – which, in turn, causes an action. Hence, the activity of the agent explains the causal background behind an action. Going back to Steps 1 and 2 of action from the last section, it is also here we can see the difference between *STEP 1* and *STEP 2* – in the latter case the capacity has been made manifest, while it has not in the former.

Is this a good solution to the problem of the disappearing agent? It differs from the solutions the literature in interesting ways. The standard solutions are based on treating certain kinds of desires as special, and hence contributing with some active force of the agent's to the initiation of actions. Standard views suggest either that the special (kinds of) desires have to be endorsed by the agent, or that some particular desire (or other mental state) plays a special role in the production of action (cf. Mayr, 2011, ch. 3-4; Schlosser, 2010). The former set of views tend to be inspired by Harry Frankfurt-style higher order-desire endorsements, according to which actions are attributable to agents if they endorse their lower-order desires with their higher-order desires (cf. Frankfurt, 1971). Significant versions of the latter view, on the other hand, include those of Velleman (1992), who adds a desire to act for one's strongest reasons as partially constitutive of the agent, and Bratman (2000), who holds that one must treat a desire as reason-giving in virtue of a long-lasting background policy to produce the right kind of action.

However, all such views are highly contentious (cf. e.g. Mayr, 2011, ch. 3-4; Henning, 2018, pp. 202-215, and references therein). But I will not present the problems for the standard views at much length here. It is enough to indicate where some of their key problems lie, and then show how my rationality-based solution to the problem of the disappearing agent above improves on them. The key virtue of this solution is that it does not appeal to special desires. In fact, *any* problem stemming from the possibility of acting on the wrong desires can be solved by my solution, since it is not based on adding desires, but on going through a certain procedure prior to acting.

First, the key problem for endorsement views is that it always seems possible to act, and even act paradigmatically, from non-endorsed desires. Whatever one's second-order desires (or other factor whereby one endorses first-order desires) might be, one can form belief/desire-pairs based on desires that one does not endorse. But this problem is easy to solve on my view. One must go through the relevant procedure, involving instrumental rationality, to form belief/desire-pairs to act. It says nothing about which desires need be involved.

Regarding the second major type of solution to the problem of the disappearing agent, it seems possible to act (paradigmatically) quite independently of the mental states that usually are supposed to play special roles. For example, one can act on one's reasons without a desire to act on them, or without a desire one treats as a reason in virtue of a policy. But instrumental rationality is less escapable than that. It is necessary for forming *any* instrumental desires.

In fact, even more recent solutions to the problem of the disappearing agent than the endorsement- or special role solutions have problems explaining how the agent is involved in the action they are supposed to issue in. Schlosser (2010) has suggested that agents 'own' their actions by default, just in virtue of having caused them. And Sinhababu (2017, ch. 10) suggests that, as the Humean self is constituted by the desires on which it also acts, all actions stem from the self by fiat. But none of them say anything about what the agent actively does to bring her actions about, whereas my suggestion is that the agent manifests her instrumental rationality when bringing actions about.

However, problems still remain for my solution. One might worry that the argument is too weak to capture the agent behind the action, or at least does not require a theory of agency that is substantive enough to solve the real problem of the disappearing agent. I have worked with a minimal version of the problem, but Korsgaard and others do not; they think it is pre-theoretical that agency involves more substantive properties than those I have discussed. For example, my formulation of the problem does not link agency to personal identity in virtue of normative features that agency supposedly has, like Korsgaard does. Korsgaard's premises (especially (2) and (3) in chapter 2, section 2) are defended by claiming that we must identify with the principles behind the action, but I have avoided appealing to identification.

However, a connection between agency and personhood can still be maintained on my view, even though it is not a case of identification. Both versions of *PARADIGMATIC AGENCY* are formulated in terms of personhood; a person is an agent in virtue of having a Humean psychology. So there is still a connection between personhood and agency here.

More generally, however, I think agency comes for rather cheap. Conceptually, any action is performed by an agent, quite independently of what agency might have to do with personal identity or other such substantive properties. The kind of agent I need is minimal; it is a collection of the properties that feature in explanations of action, whichever they are, and a person has. And the minimality of this conception of agency is a strength. I need not worry about how the self or the person or their unity might work, let alone personal identity over time, while still having something to say about paradigmatic actions.

Another worry is that I might only have explained how instrumental rationality is constitutive of *some* actions, and hence only solved the problem of the disappearing agent for a subset of all actions. I have even explicitly formulated my understanding of 'paradigmatic actions' so that paradigmatic actions are successful actions. But what do I make of unsuccessful actions? Furthermore, I have not ruled out the possibility that there are some actions that do not have a means-ends structure, or possibly even non-paradigmatic agents (who only are able to act non-paradigmatically). What do I make of such cases?

Actions performed by paradigmatic agents, whether these are successful actions or not, can all be treated in the same way. In all these cases, the agent still has the principle of instrumental rationality psychologically available, and hence still stands behind the action when performing it. Admittedly, the agent may exercise the principle to a lesser extent, or not at all, in some cases of non-paradigmatic actions. But it is not strange that the agent cannot actively initiate non-paradigmatic actions in the same way as paradigmatic actions – these actions are plausibly non-paradigmatic in part *because* they do not involve actively exercising the capacities that we typically exercise when acting (cf. section 7 below). We can even generalize the last point to also explain why non-paradigmatic agents are non-paradigmatic: they do not actively exercise capacities or stand behind their actions to the same extent as paradigmatic agents, and are for that reason less paradigmatic.

A final worry is that the agent might not seem to be active *enough* on my solution to the problem of the disappearing agent. I did claim that the agent has to be active in the initiation of action, after all, and one might argue that the kind of activity involved here is not enough for paradigmatic action, perhaps because the manifestation of a capacity is too hampered by being part of an event-causal order. Or maybe activity demands something stronger, such as first-personal self-awareness. I have not required that here, but I have argued that this might be a feature of agential activity in the past (Leffler, 2016).

The former worry should be fairly easy to assuage, however. A manifestation of a capacity on part of an agent is a kind of activity and differs from the non-manifestation of the capacity by displaying itself. It can succeed or fail in doing that, so it is not the same thing as having a property that cannot either succeed or fail to manifest. Hence, a procedure involving a capacity is active in a way that a non-manifesting procedure is not.

The latter worry is a bit more complicated. It may well be that agential activity or control, fully characterized, involves first-personal awareness. But simultaneously, we probably do not need to be active in all senses to perform a paradigmatic action (as opposed to, e.g. the most sophisticated action possible). Requiring first-personal awareness of the procedure seems overly intellectualistic in pedestrian action cases (cf.

the discussion of Velleman's view in chapter 2, section 3). And my kind of agent is active enough to perform paradigmatic actions.

(5) Normative Force

By now, I have argued that instrumental rationality is constitutive of paradigmatic agency, most importantly because it helps solving the problem of the disappearing agent. The second major aim of this chapter is to show that instrumental rationality has normative force (or 'prescriptivity' or 'normativity'). Some think that this is a mistaken question to ask (Levy, 2018), but for all I have said so far, the instrumental principle may still fail to be normative – its normativity has often been challenged (cf. e.g. Kolodny, 2005; Korsgaard, 1997; Raz, 2005*a*; 2005*b*). Moreover, the explanation of its normativity might stem from some source that would threaten the success of the constitutivist project, e.g. some non-reductive non-naturalist one (cf. Hussain, ms.). So an explanation of its normativity – and the right kind of explanation – must be given.

I shall now take a shot at providing that explanation. To do so, I shall first characterize the kind of normativity it involves (here in section 5), and then explain how instrumental rationality has it (in section 6). This positive explanatory strategy will, I hope, be able to bypass various objections just in virtue of being such a strategy.[8]

However, to explain the normativity of instrumental rationality in the right way, I first need a theory about what it is to be normative, in the relevant sense.[9] There are different conceptions of normativity in the literature, so I need to settle on some view. In recent discussions about the normativity of rationality, most have thought that for *x* to be normative, *x* has to be reason-providing, and is normative because it is (cf. e.g. Kiesewetter, 2017, ch. 1-2; cf. Wedgwood, 2017, ch. 4 for criticism).

But there are also other possible views, such as (broadly speaking) value-based ones. For example, some think that some forms of normativity can be explained in terms

[8] For example, Wedgwood (2017, ch. 8) argues that constitutivists cannot explain the normativity of rationality because it is implausible that we always act according to constitutive norms, and if we can avoid that, we can be shmagents. (Nor does he think propositional attitude-based constitutivism works.) But part of my story will be to show how we can fail to act in a way that is still governed by instrumental rationality. Enoch (2011*b*; cf. chapter 3, section 2) thinks that explanations in terms of inescapability do not explain normativity, yielding a naturalist fallacy. But he has not discussed plight inescapability, just dialectical inescapability, and, together with other properties, plight inescapability can do serious work here. Moreover, many philosophers, e.g. Kiesewetter (2017, ch. 1-2), assume that the normativity of instrumental rationality has to be one of reasons, and then argue that it does not give reasons. But I will present an interpretation of the normativity of rationality that has nothing to do with reasons.

[9] Cf. Finlay (2019)'s property 'normativity$_{\text{ont}}$', i.e. the property of normativity, rather than a normative property.

of how much value they promote, given that states of affairs have value already (e.g. Moore, 1903). A different value-based view is attributivism, according to which members of some kind can be ranked by how well they live up to the function of that kind, and are good or bad to the extent that they do so (e.g. Thomson, 2008; Smith, 2017). Or it might be that rationality fundamentally is a virtue concept (cf. Wedgwood, 2017).

It is unclear to me why rationality would be value-based, as the concept seems deontic. But instead of developing that critical point in depth, I aim to provide a positive argument by delineating a form of normativity which is apt for my purpose, and then explain how my principle of instrumental rationality has the right kind of normativity. My strategy, then, will be to pick out some – hopefully all – of the properties of the normative force that rational principles plausibly have, and show how they can be explained by inescapable paradigmatic agency. The latter, then, will be the source of normativity of rationality. To the extent I can do that, I will be showing that instrumental rationality is *structurally* normative; it features in the structure of (inescapable) agency and has the relevant properties to be forceful just in virtue of that.[10]

But what is normativity? Just what the notions of 'normativity' or 'prescriptivity' mean is unclear, but the core intuition is that there is something about normativity or prescriptivity which tells agents what to do, and it – the 'oomph' or 'force' – can be captured by verbs such as 'requires', 'recommends', or 'prescribes.' The normative oomph or force of X is the extent to which that thing points agents in some direction that X requires, recommends or prescribes them to take.

Little work has, however, been done to give a comprehensive account of the features of normative force that we usually associate with rationality. In line with the properties I suggested we explain in section 1 above, it seems like an explanation of the normativity of rationality must explain a specific set of properties. It must show that the principle can have a grip on us independently of what we feel/want/desire to do at any given moment in time, for it applies to all paradigmatic actions, and of how it, phenomenologically, appears to be normatively constraining. Moreover, the normativity of *IR* should, in virtue of stemming from a principle, not be aggregative, in the sense that

[10] There is therefore a sense in which this explanation appears reductive. It only appeals to features of agency that seem to be reducible to descriptive properties. Jean Hampton once called this a kind of 'psycho-social authority' (Hampton, 1998, p. 99), and contrasted it with 'objective authority', which she took be more genuine. Many, like her, will want to deny my kind of project. However, first, I am not fundamentally committed to the reducibility of the mental, so I am not a whole-hearted reductionist (cf. chapter 8, section 4). Second, even so, insofar as we can provide a reductive explanation of normativity, it is preferable to otherwise problematic explanations that require more.

IR counts as a consideration among others in deliberation. It is a structural norm, not a reason-giving one.

There are also some more general features of normativity that presumably are true for a principle of rationality. One should be able to show how we can be normatively criticizable for failing to live up to the principle, ideally *in virtue of* failing to live up to it – and how one can be criticizable for failing to live up to it independently of whether one gets one's reasons for actions right. Moreover, one should both be able to live up to the principle and to fail to live up to it. A theory that is able to explain all these features about the normativity of rationality seems *prima facie* justifiable.[11]

To pre-empt my conclusions, I shall try to explain the normativity of instrumental rationality because it sets a directive standard of success (cf. Copp, 1995, ch. 2; Katsafanas, 2013, ch. 2). The core idea here is that one can either live up to, or fail to live up to, some *standard*, hence making the standard one of success. A standard specifies a set of conditions to be lived up to for things of a certain category, and that it is possible to fail to live up to (cf. Copp, 1995, p. 19).

Such conditions can have many sources; for example, they are often, but not necessarily, set up just by what one is aiming to do. If 4-year old Xenia is trying to tie her shoelaces, she succeeds if she does so and fails if she does not, relative to her aim of tying her shoes. But standards of success can also depend on other things than aims. There are grammatical rules in languages, and speaking some language requires living up to its rules well enough whether one tries to or not, on pain of incomprehensibility. Even such standards of success can give us one kind of normativity.

But a standard is not, by itself, enough here. Even though standards of success can tell agents what they need to do to live up to them, they do not by themselves have normative oomph or prescriptive force. If Xenia is not trying to tie her shoelaces, there is no normative pressure on her to do so (*ceteris paribus*, at least).

[11] The other most worked out attempt I have seen that lists what the makers of normativity for principles of rationality is that of Southwood (2018*b*). Citing Schroeder (2011), Southwood lists the following features: (*i*) such principles should bring about deliberative closure, in the sense that they settle what one is to do, (*ii*) they should matter for how we give advice, (*iii*) they imply that we can do what they require, and (*iv*) they can be grounds for criticism. Moreover, (*v*) they must be closely – maybe conceptually – connected to obligation, though Southwood denies that this is true for rational normativity.

Southwood's (*i*) seems very much like my point about how rationality can have a grip on us independently of what we want/feel/desire to do at any particular point in time, and his (*iii*) and (*iv*) are general conditions of normativity that I accept. However, we can probably ignore explaining how normativity matters for the purpose of giving advice, because presumably something matters for our giving advice *because* it is normative. And I share Southwood's scepticism when it comes to thinking that the connection to obligation matters. We do not ordinarily take ourselves to be *obligated* to be rational.

We need something stronger, and I shall call this property *directivity* (cf. Hampton, 1998, ch. 3). For something to be directive, it must prescribe something that agents should do by pointing them in the direction of that thing. So, for example, if a road sign features an arrow pointing left, it directs drivers to steer their cars in that direction. But even more is needed for directivity in my sense. One must capture the pre-theoretical description of normativity, i.e. independence of present desire, feeling or want plus the appearance of constraint, sharp edges in virtue of being the normativity of a principle, the ground for criticizability (independently of reasons), as well as normativity-implies-can and can-fail. How do we get all that?

(6) Explaining the Normativity of Instrumental Rationality

To start off, I shall explain how IR_{HUMEAN} is directive independently of what one aims to do or may feel/want/desire at any particular moment in time, and hence applies in all situations. I follow Korsgaard (2009, pp. 1-2) in thinking that the principle of instrumental rationality is inescapable because following it is constitutive of agency, and action and agency are our *plights* (cf. chapter 2, section 2; chapter 3, section 2). We are continuously faced with choice situations where we must make up our minds, not least by exercising our instrumental rationality, forming instrumental desires by which we may act. We may not knowingly try to do so, but we are still confronted with situations in which we must be sensitive to combinations of mental states, and hence cause instrumental desires by which we act. Not least, intentionally omitting to act, or trying to end one's agency, both similarly involve making use of the principle.[12] Hence, there is a practical transcendental argument which shows the inescapability of *IR* – to be an agent of the kind that the world makes us be requires instrumental rationality, and we cannot avoid agency, so we cannot avoid instrumental rationality.

Moreover, importantly, the situations that we face where we can exercise our agency are such that we are forced to use *all* the capacities of our agency in them as well as do so fully. This is because that is what we inescapably face doing. Hence, plight

[12] In fact, more specifically, plight inescapability encompasses the continuous new situations we face and dialectical inescapability, since acting so as to stop acting is a special case of action. Dialectical inescapability, the reader may remember, stems from two properties: the fact that agency is the enterprise of the largest jurisdiction, and that it is closed under reflection (cf. chapter 3, section 2). The latter is true *because* we are continuously faced with new choice situations.

inescapability captures how acting *paradigmatically* is our plight, not how acting in any possible way is.

Plight inescapability does not say that agency in inescapable in all possible ways it might be inescapable, however (cf. Ferrero, 2018; 2019; cf. chapter 3, section 1). We can avoid agency by (e.g.) dying, or by acting in ways that use less than all our capacities, including the capacity for instrumental rationality. But it explains that we always are in a position where we *face* doing so, given the choice situations we face. Paradigmatic action is our plight because the world, inescapably, forces us to act paradigmatically.

Plight inescapability can positively contribute to our explanation of normativity. As we are continuously faced with choice situations in which we need to make up our minds, we are forced to manifest our agency, insofar as we have the features that constitute *PARADIGMATIC AGENCY*$_{IR}$. And these features apply every time we are forced to make up our minds, for paradigmatic action is plight inescapable, not just any old action. This establishes a standard of success for what we do regardless of what we do, for we can either successfully manifest the features of her agency or not.

Now, to explain normativity, I need to show how the standard of success inherent in agency applies independently of what one happens to desire, want, or feel. Assume, then, that someone is an agent, so she has beliefs, desires and the instrumental principle, and she does not act. Then she is continuously faced with new choice situations, which she cannot avoid. And she is confronted with such situations all the time, and any particular action she may perform will *ipso facto* have her acting. So regardless of what she attempts to do, she must manifest the features of her agency.

Hence, the plight inescapability of *IR*$_{HUMEAN}$ can explain that it holds regardless of what one may feel, desire, or want at any given moment. Regardless of what one chooses to do in ordinary choice situations, one must manifest one's agency (and do so fully). If so, it seems like one must manifest it regardless of what one does. So insofar as one acts in the standard way in which one is forced to act by plight inescapability, one must adhere to the principle of instrumental rationality.

I have now shown why instrumental rationality is a standard of success regardless of what agents may feel, want, or desire. To explain why it phenomenologically seems to have force by appearing constraining, I shall introduce a number of properties inspired by some properties of (intentional or personal) commitments, and show that it has them. I do not, however, mean to say that commitments themselves are normative, and normativity need not be understood in terms of commitments – rather, the point is that the phenomenology of commitments is *analogous* to that of normativity. The relevant properties are:

> (*GOAL*) Commitments are intentional (in Brentano's sense), and hence there is something they are about. So they *have a goal or an aim*, which is their object, and they help one achieve (Chartier, 2018, ch. 2).
>
> (*ENGAGING*) Commitments guide one towards doing something, in a specific way (Bratman, 1999, ch. 7).
>
> (*SYNCHRONIC OR DIACHRONIC*) Commitments can be *synchronic or diachronic* (Chang, 2013). One can have one just for a particular moment, or they can extend in time. In at least the latter case, they provide structure or direction for one's life (Chartier, 2018, ch. 2).
>
> (*SETTLING*) Commitments bear an interesting relation to *settling* courses of action (Bratman, 1999; Chang, 2009; 2013; Chartier, 2018, ch. 3). *Ceteris paribus* (i.e. absent other, stronger, commitments or norms), a principle or commitment settles what one is to do, hence determining one's future.
>
> (*RESIST RECONSIDERATION*) Finally, commitments *resist reconsideration* (Calhoun, 2009). They tend to crowd out re-opening deliberation regarding whether one is to pursue their object (or form other intentions), though again *ceteris paribus* (e.g. absent significantly altered life circumstances or desire sets).

These properties provide an appearance of normative constraint. In particular, commitments set aims that one engages with (cf. *GOAL* and *ENGAGING*), settle what one is to do (cf. *SETTLING*), and cannot be very easily altered (cf. *RESIST RECONSIDERATION*). And it is this appearance of constraint that we are after capturing, even for principles. Importantly, whatever has them – commitments or principles – stands in contrast with desires. Chartier (2018, p. 1) even *defines* commitments in part as being such that they cannot be altered at will. Whether they can be defined like that or not, desires at least lack *SETTLING* and *RESIST RECONSIDERATION*.[13]

[13] Moreover, weaker forms of normativity than that of instrumental rationality may have some but not all these properties. Plausibly, Xenia's shoe-tying may have *GOAL*, *SYNCHRONIC OR DIACHRONIC*, and *SETTLING*, but fail to engage her in the action or make her resist reconsideration – that is part of the reason she may just quit tying her shoes. This means that other norms can seem constraining to various degrees, which seems like the right result: norms can be experienced as more or less forceful. This means that the standard distinction between standard-of-success-based normativity, which need not be forceful, and more comprehensive forms of norms to some extent is artificial (cf. Finlay, 2019).

IR_{HUMEAN} captures these constraining-seeming properties, showing how the principle appears to be more constraining than any given desire. It captures *GOAL* because agents have aims just in virtue of desiring, and need to take means to satisfy them on pain of irrationality according to the principle. It is *ENGAGING*; it makes agents *engage* with some goal, for it is what links their mental states up so they can act to attain it. It explains *SYNCHRONIC OR DIACHRONIC* because agents can perform both instantaneous actions and actions that extend over time.

Moreover, IR_{HUMEAN} captures *SETTLING*, because once one has formed an instrumental desire in virtue of it, that desire will make one act (*ceteris paribus*, at least). And it captures *RESIST RECONSIDERATION* because once one has settled on acting on some desire (and is otherwise functional), one will do so unless something important comes up to make one reconsider. The qualification here gives us the right result. One should be able to change the desire one acts on, but not completely arbitrarily.

So IR_{HUMEAN} captures want/feel/desire-independence and the appearance of constraint that normative principles have. What about the other properties of normativity? First, I claimed that the normativity of IR_{HUMEAN} is not supposed to be aggregative; IR_{HUMEAN} is a structural principle, not a reason, desire, or value. That is easy to capture here. IR_{HUMEAN} comes in the form of a principle, not a reason, desire, or value. One can succeed or fail to live up to it, but it would be a category mistake to add it up with reasons, desires, or values. Its role lies deeper down in the causal genesis of action; it plays a coherence-inducing role in the formation of instrumental desires, not a goal-setting role like desires (or reasons or values).

Second, there are also some more general features of normativity that I need to capture. These are reasons-independent criticizability in virtue of failure, ought-implies-can, and ought-implies-can-fail. We can explain criticizability in virtue of how we may fail to exercise our capacity for instrumental rationality, for when we fail to do that, we fail to manifest our agency behind our actions. Whatever we do without it fails to be an action we author fully, and we do not act from the most fundamental kind of action, which is successful action (cf. section 7 below). That failure is criticizable, since we are not utilizing our capacity for instrumental rationality, and hence not living up to the universal standard of success that it sets. And this criticism applies to agents even if they do not respond to their reasons, for one can act on a desire for the bad but still fail to take the means to it.

Finally, we can both live up to, or fail to live up to, the principle. It is usually not hard to take the best means, in the weak sense where it is ‘the thing to do’ in any given situation (cf. section 2). And we can, similarly, fail to take the best means (though cf. section 7 below for complications about failure).

Instrumental rationality, then, features in our paradigmatic actions as an inescapable standard of success that also captures the properties I have associated with normativity. This lets us answer the question of whether rationality is normative. We can say that means-ends rationality is structurally normative, insofar as the way that it features in actions explains how it is normative.

Are there possible objections? Of course. One might think that this argument goes by a bit too fast. Perhaps there is some weak sense of action in which one has to do something all the time, or at least most of the time, but that does not show that one always has to have linked-up beliefs and desires. Yet that is what I just claimed.

The argument assumes, however, that the person (or creature) acting already is an agent. Someone who has yet to develop beliefs and desires (or the ability to unify them by linking them up) is not a paradigmatic agent, so we can ignore people (or creatures) who are not. It is plausible, for example, that even though the Martians or Saturnians of chapter 3 have reasons, they need not have this principle, because their psychologies are intrinsically different from ours. And it is plausible, I claimed in section 4, that we can capture which agents are less paradigmatic than standard humans in virtue of seeing how they fail to live up to the principle of instrumental rationality.

Is there a way in which someone who is an agent could avoid acting paradigmatically? Well, she could decide not to act. But doing so would have her manifesting her agency. She would, self-defeatingly, be acting against acting, as per dialectical inescapability. And that action would have her committed to everything in the instrumental principle in the first place, since, being a paradigmatic action, it would involve instrumental rationality.

What about just *omitting* to act? Some think that omitting to act in fact is not to act (e.g. Clarke, 2010). This distinction has even been used to defend causal theories of action like *HTM* – in the paper just cited, Clarke argues that such theories can be true about actions, though they need not be true about omissions. Could one not, then, *omit* acting, and hence not act? No, because omissions are actions, on my view. In appendix A, I argue that instrumental desires can cause events by *causally sustaining* them. Omitting is to decide to, or at least make oneself, causally sustain the present state of the world. That is to act, in a broad sense of the word.

Alternatively, could agents avoid acting without deciding not to act? It is possible that their beliefs or desires need not be sufficiently linked up to act. They are agents because they have beliefs, desires and the instrumental principle, but they may perhaps not manifest the principle at any particular point in time.

The answer is no, in the broad sense of acting that is at work here. For insofar as their mental states still cause something (whether they initiate it or sustain it), what they do counts as an action. The action in question is more often than not a non-paradigmatic action. But that does not mean that they are not acting. Moreover, there can still be cases where agency is diminished. Perhaps some agents cannot sustain or initiate any causal chains in virtue of their mental states; maybe they are comatose. Then the agents do not act. But, in such cases, they are wholly outside agency.

Is this still too hasty? There are two ways in which actions can avoid living up to the constitutive standard of paradigmatic action. Some think that there are other cases of action, or even agency. First, some actions – e.g. actions out of emotions, habit, or skilful actions – might lack a means-ends structure. However, whether or not such cases indeed are actions, they cannot be sustained for long. Since the agent is continuously faced with new choice situations in which they can take means to their ends, only a few fleeting moments of such actions seem possible for paradigmatic agents. So these cases are not problematic. They are slip-ups.

Second, some actions can be failed attempts to live up to the constitutive aims of action. Such failures are still possible to assess via the standard of action. But then they are no longer paradigmatic actions; they are flawed versions of standard actions, implicated in the same structures. That is what I shall argue in the next section. This means that an agent performing failed actions still has all the properties of action and agency, which is sufficient for necessitating action for them. And then we can stick with the explanation above.

(7) The Problem of Bad Action

I have indicated that I need to explain how one can fail to adhere to the constitutive standards of agency, and indeed also that I can provide such an explanation, as I have claimed that agents can fail to take the best means to their ends. However, explaining how that can be the case is a very general problem for constitutivism, known as *the problem of bad action* (Clark, 2001; Lavin, 2004). This is the problem of explaining how one can act and fail to follow some norm if following that norm is constitutive of action. For if the norm is constitutive of action and one fails to follow it, one does not seem to be acting anymore – and, *mutatis mutandis*, the same problem emerges for other possibly normatively constituted aspects of agency.[14] In the Humean case: if it is constitutive of paradigmatic action (or agency) to take the best means to one's end, how can one act *and fail* to do so?

Even worse, this problem seems particularly pertinent for constitutivism about instrumental rationality. The problem of bad action has been employed to argue that instrumental rationality by itself cannot be normative because there is no way to do wrong on a Humean picture of agency (Korsgaard, 1997; cf. Hampton, 1998). Korsgaard argues that the prescriptivity of instrumental rationality presupposes something stronger – a categorical imperative, perhaps. The problem is that it is possible to fail when acting, but Humeans are committed to saying that one's strongest desires cause one's actions, and taking the means to them is constitutive of action. So bad action is impossible, so Humeanism fails.

Prescriptivity understood in terms of a directive standard of success should make it abundantly clear that one can fail to live up to the instrumental principle, however. Perhaps this is because, as Sinhababu (2011) has emphasized, one momentarily focuses one's mind on some other desire than the one which is one's end, or at least another end than the one that one is attempting to reach (cf. Hubin, 2001). For example, when I have been writing this dissertation, I have sometimes drifted off to procrastinate by watching Youtube videos. It seems wrong to say that I have done so because of my end-desire, even though the desire to watch Youtube videos has been more strongly occurrent for me at some points. So we can explain at least one type of irrationality.

There are more significant cases of bad action, however. Most importantly, one can be akratic (cf. Davidson, 1980*c*, who stresses this against Hempel, and Wallace, 2001; 2004; 2013, who stresses it against Korsgaard). Perhaps one takes the wrong means to one's end, perhaps one does not form the right instrumental desire to attain that end, or perhaps both, so one tries to take a bad means to a desire which is not one's end. Sinhababu's case instantiates the first form of *akrasia*, but clearly there are also other forms of it.[15]

[14] Sometimes, this point is also made by saying that constitutivists need to explain why someone can fail to be an agent at *some* times without therefore having their agency undermined (cf. Wedgwood, 2017; Lord, 2018).

[15] I do not mean to say that these three sorts of *akrasia* are all the sorts there are, however. As they are forms of *akrasia* present in the formation of instrumental desires, we may call them formation-*akrasia*, but there is also execution-*akrasia*, where one fails to execute one's best normative judgment and instead goes

The standard constitutivist solution to this kind of problem is to hold a (so-called) threshold view. Here, acting only requires living up to a norm in a minimal sense – metaphorically, in a sense that passes a threshold (Lindeman, 2017; cf. e.g. Korsgaard, 2009, ch. 2; 5; 8). Seeming actions that do not live up to the norm at all do not count as actions. By analogy: a house without walls may not be a house, at all, but this leaves room for poorly built houses to still count as houses. So there can be poor actions, where poor actions involve living up to the function of action (or manifesting agential capacities) less than fully. By contrast, seeming actions that do not at all involve the function of actions do not count as actions at all on this view.

This explanation of failures *might* work.[16] In the case of instrumental rationality, someone who takes necessary means to a desire which is not an end-desire, or who fails to take the necessary means to some desire which is an end-desire, and hence fails to live up to it, may in some sense be less than fully instrumentally rational.

But the third type of *akrasia* I mentioned poses a problem here. We can exemplify it with a procrastination case which seem highly rationally problematic. The idea is that someone who fails to take the necessary means, but still takes *some* means, and does so to a problematic end, still acts. If my end is writing this dissertation, I (akratically) desire to watch Youtube videos, but I start looking at Facebook instead of Youtube when procrastinating, then what on earth am I doing? At least one interpretation of what is going on in such a case is that I fail with respect to *all* the components of the norm of instrumental rationality. I am not acting on my end, and I am not taking the right means. Accordingly, on at least one interpretation, what I am doing seems too deficient to involve the function of action or successful manifestations of my capacities in *any* sense.

There is a quick and dirty solution available for threshold theorists. If action or agency come in degrees, one can say that such cases still are actions, but actions where the constitutive principle of instrumental rationality is actualized to the degree 0 per cent – and hence not at all. One could then still treat them as, in a way, belonging to the same kind of things as ordinary actions. The quickness and dirtiness of this response raise more questions, however. First, the move looks *ad hoc*. In the Youtube over Facebook case, I

for some other one. Just how execution-*akrasia* is to be explained is an interesting question, but it is separate from the explanation of bad action.

[16] Or, at least, it might make sense in the context of instrumental rationality, which is a fairly weak norm. Korsgaard (2009, ch. 8) gets into big trouble when appealing to this solution, for she cannot really explain how someone can seem simultaneously evil and competent. But Korsgaard's general discussion of bad action is confusing and even contradictory. At times, she requires *CI* to be the constitutive principle of agency, and at times, agents are allowed to be constituted by other principles (cf. Bachman, 2018). It is best to stick with a solution that is relevant for present purposes instead.

am plausibly acting, but I do not seem to be manifesting my capacity for instrumental rationality at all. Second, it is not clear where the scale of actionhood comes from. What counts as an action, and what does not, here?

It might be possible to answer these questions. But I would, ideally, prefer to go with an alternative solution. Constitutivists have proposed a number of solutions to the problem of bad action, and I am most attracted to one that allows for both failed and successful cases of action.[17] To emphasize this, we might label the solution I have in mind a kind of *disjunctivism*. It allows us to say that there can be failed cases of action that do not include the active manifestation of the instrumental principle at all, as well as cases of deficient manifestation.

How does it work? Action disjunctivism comes in some different guises and has been used to solve some different problems (cf. Lord, 2018, pt. III). In perhaps the most familiar version, the disjuncts have had to do with reasons. According to Hornsby (2008), an agent either acts from a reason she knows she has (and hence represents accurately), or from a reason she merely believes she has. My version is also epistemologically inspired, but in another way. In Timothy Williamson's epistemology, knowledge is the most fundamental factive mental state. To believe that *p*, however, is to treat *p* as if one knows it, whether one does so or not. This means that non-factive beliefs are failed knowledge, though factive beliefs *may* map what we know (Williamson, 2001, ch. 1).

This view serves as my inspiration. For, analogously with how Williamson treats knowledge, one may think of successful action as the most fundamental action explanation, and then treat cases of successful action as the most fundamental *kind* of action.[18] Then one can say that action fundamentally constitutively involves an event non-deviantly caused by beliefs and desires rightly combined by the instrumental principle, as it does in successful cases, *or* various deviations from that fundamental kind of action. These deviations are either failed actions or actions involving exercising one's agential capacities to a lesser extent than actions of the fundamental kind, and may or may not be

[17] Alm (2011) has a constitutivist view with a structure which is fairly similar to the one I shall propose, but instead of talking about kinds of action, he talks about 'defaults'. Two other recent suggestions – from Pauer-Studer (2018) and Dick (2017) – are unhelpful for present purposes, however. Pauer-Studer develops a second-personal form of constitutivism, but I have noted to discuss such views in depth (cf. chapter 2, section 1; chapter 7, section 4). And Dick, drawing on Bishop Butler, argues that an action counts as an action if it stems from the right kind of creature, but leaves it open to us to be that creature or not. But instrumental rationality is not optional in that sense.

[18] My talk of 'fundamental kinds' here comes from another disjunctivist: Martin (2004) appeals to a 'fundamental kind', which provides the deepest answer to questions about what perception essentially is. This involves veridical perception. Then there are other, non-fundamental, instances of perception (e.g. non-veridical perception), which belong to some less fundamental kind(s).

what one would have done had one acted successfully. And, importantly, the latter disjuncts are non-fundamental kinds of action. However, the fundamental kind of action has primacy over the latter cases because they are failed actions or actions that make lesser use of one's agential capacities.

There are several examples of actions that belong to non-fundamental kinds. Cases of outright failure to reach one's ends notwithstanding, there might be means-ends actions featuring a desire and a belief that would have caused paradigmatic actions if they had been linked up by IR_{HUMEAN} and non-deviantly caused an event, but now (non-deviantly) cause some event even though the desire and the belief have not been linked up. Moreover, *if* there are such actions, we may treat cases of action that do not involve beliefs, such as actions out of desires that do not involve taking means to ends, as belonging to non-fundamental kinds. In all these cases, one acts from less than beliefs and desires linked up by the instrumental principle, and the actions performed may not involve taking means to one's ends, for the mental states explaining them are not successfully linked up by the principle. However, importantly, actions without the instrumental principle or even without beliefs *may* conform to what one would do if one were to act successfully, for one may acquire the same results by such actions as one would have done if one had acted paradigmatically.

On this view, we can allow for actions without even the minimal amount of agential functioning or capacity manifestation one needs to act on threshold views. But such actions are still *governed* by the norm of *IR*, as they belong to non-fundamental kinds. The fundamental kind of action (and agency), however, is one that will involve at least beliefs, desires, and a link between them via the instrumental principle. And this is because of the argument in sections 3 and 4 above. The instrumental principle is needed to solve the problem of the disappearing agent.

We may then use the recently introduced action disjunctivism to explain all the akratic cases I have mentioned, so we can avoid the awkward solution to akratic cases that threshold theorists have gone for.[19] In all successful cases, (paradigmatic) actions involve beliefs, desires, and successful linkage by the instrumental principle. But one can fail in all relevant ways here, generating cases of irrational actions.

[19] In fact, there is also another case that can be avoided. In conversation, several philosophers have stressed cases of faulty instrumental desire-formation, e.g. cases of deviant causation. Such cases are clearly cases of faulty action even on the present view, however, since there is a success condition involved in paradigmatic action. Paradigmatic action involves having formed instrumental desires correctly.

We can see how this works by returning to the Youtube case. If I go on Youtube instead of keeping on working, I fail in one way, though I may take the necessary means to do so. Alternatively, if I try to work but keep thinking about some Youtube video I have watched rather than writing, I am not taking the necessary means to my end. Or, if I try to go on Youtube but type the URL of Facebook instead of Youtube into Google, I fail with respect to both means and ends. I have the wrong end (Youtube rather than working) but I take the wrong means even to that (by typing in the Facebook URL). All these cases involve failure of some sort, and in the same way. For had I applied the principle of instrumental rationality correctly to my desires and beliefs, I would not have acted in the way that I do. But I do, and hence display various forms of *akrasia*.

Importantly, this solution to the final Facebook URL case differs from the solution that a threshold theorist might provide. A threshold theorist would have to insist that the case should be interpreted with the principle at least minimally active (over some threshold). This is what led to the 0 per cent action problem above. But the disjunctivist view can handle a version of it even when the principle is not active at all. It can generate an action even though not all the components of action are involved in the particular case. The disjunctivist can therefore say that while all paradigmatic actions are governed by instrumental rationality, some actions need not involve exercising it at all.

Moreover, the disjunctivist solution here is not *ad hoc* or unprincipled like a threshold response might be. It is not an *ad hoc* extension of the view as much as it shows how to explain the deficiency of deficient cases; it explains the relevant kinds of *akrasia* by saying that such actions deviate from the paradigmatic case. So we can make sense of bad actions in a theoretically laudable manner. Furthermore, we can count any actions, whichever they are, that fail to use all the capacities of Humean agency, or fail to do so properly, as non-fundamental kinds. This means that even though the view is contingent with respect to how many kinds of action there are – I have not taken a principled stand on whether there are, for example, actions out of desires that do not involve beliefs – I can still explain the relation between different kinds of actions in a principled way.

(8) Conclusion

This has been a long chapter. In section 1, I introduced the idea of constitutivism about instrumental rationality. In section 2, I characterized the instrumental principle. In section 3, I argued that Humeans should accept it to show how the agent stands behind her actions and actively initiates them, and in section 4, I argued that this move improves on the standard solutions to the problem of the disappearing agent. Then I discussed how this principle is normative: I explicated what normativity might require in section 5, showed how instrumental rationality is structurally normative in section 6, and solved the problem of bad action in section 7.

In section 1, moreover, I introduced several *desiderata* for a successful form of constitutivism about instrumental rationality (given that it is supposed to do work in my theoretical project). The explanation needs to (*a*) capture the pre-theoretical characterization of *IR* – i.e. it has the form of a principle which is normatively forceful because it applies to all actions and feels more constraining that arbitrary desires, feelings or wants, does not aggregate, can ground criticism if we fail to live up to it (where that criticism is independent of our responses to reasons), and is such that we can, or can fail to, live up to it. Moreover, the explanation must (*b*) be able to be part of an explanation of the normativity of reasons, and it must make sense of the *desiderata* (*i*) and (*ii*) from the shmagency discussion – i.e. it should be able to explain reasons for shmagents, and be able to avoid underdetermination worries.

Can IR_{HUMEAN} do all this? I believe so. Re: (*a*), I now have a view where rationality comes in the form of a normative principle. It is plight inescapable, and it captures how normativity appears constraining. Moreover, instrumental rationality has a sharper kind of normativity than reasons. It involves successful or failed cases, so there is no question of aggregation, but we are instrumentally required to live up to the desires that set our ends. And our failures are easily criticizable because we may fail to live up to the standard of success that plight inescapability sets. Finally, we can, of course, live up to the principle – or fail to do so, as according to the solution to the problem of bad action.

Regarding (*b*), I shall try to explain the normativity of reasons in the next chapter. And regarding the *desiderata* (*i*) and (*ii*), I have presented a descriptively inescapable version of constitutivism about rationality. While this gives me the tools to develop a form of partial constitutivism about reasons, the normative phenomenon I have discussed in this chapter is rationality, not reasons, so it irrelevant that shmagents do not have it. It is plight-inescapable for persons, not for Martians or Saturnians. Moreover, the principle does not come in degrees – instead, action comes in kinds – which means that there is no question about whether or not it may end up underdetermined. It may not.

6. Desires Internalism (Probably) Explains Everything

It is now time to develop the second tier of my constitutivist framework. To do so, I shall defend a form of desire-based practical reasons internalism. The view I have in mind is inspired by Williams and Smith, and is based on the constituents of *PARADIGMATIC AGENCY$_{IR}$* – viz. beliefs, desires, and instrumental rationality, understood in a functionalist way.[1] In this chapter, I defend it in two steps. First, I argue that this type of reasons internalism can explain many key features of reasons. Then I play defence, responding to a number of well-known objections to the view. These moves pave the way for defending a take on moral reasons in chapter 7.

I proceed as follows. In section 1, I explicate my theoretical strategy. In section 2, I present a number of key features of practical reasons that any theory should explain – and immediately proceed to show how my preferred theory captures them. Then I reply to criticism. In section 3, I reply to some general challenges to reasons internalism, and in section 4, I defend an advisor version of the theory in response to the conditional fallacy challenge. I conclude in section 5.

(1) Strategical Remarks

Reasons internalists think that all reasons for action are necessarily related to an agent's motivations, actual or ideal. This is just as true regarding ambitious moral reasons (such as to stop climate change) as it is for trivial everyday reasons (such as to drink a glass of water). Reasons externalists deny this. According to externalists, at least some – though not necessarily all – reasons may lack relations to agents' motivations.

I shall defend a version of reasons internalism according to which (ideal) desires explain reasons. Because reasons depend on desires here, theories like this one are often called *Humean*. However, I want to make clear that they are logically distinct from *HTM* or Humean theories of agency (cf. chapter 4; 5) – one can hold some of these theories without holding the others. Moreover, some authors restrict the term 'Humean theory of reasons' to unassuming versions of desire-based theories about reasons, labelling more ambitious forms 'non-Humean' (cf. Smith, 1994, ch. 5). But I shall use the term inclusively, counting my view as Humean.

[1] Many others have also defended similar theories (e.g. Brandt, 1979; Goldman, 2011; Joyce, 2001; Markovits, 2014; Persson, 2013; Strandberg, 2018; 2019).

Furthermore, Humean theories of reasons need not be instrumentalist. Even though reasons are based on our desires here, they are usually theories of 'final', not instrumental, reasons. But while how instrumental reasons should be explained on Humean views is an interesting question, that question is orthogonal to the question of whether reasons should be explained by desires in general. Hence, I shall focus on final reasons and just assume that my theory generalizes to instrumental reasons.

Moreover, I shall leave epistemic reasons, aesthetic reasons, and reasons for emotions to the side, instead focusing on reasons for action (or 'practical reasons'). What are practical reasons? A conventional paraphrase – though not an analysis – is that they are considerations that count in favour of actions (cf. Scanlon, 1998, p. 17). I take that to be a necessary, but not obviously sufficient, feature of practical reasons; it is possible that other considerations do that too. But as far as practical reasons are concerned, I shall argue that they can be explained by:

> (*DESIRES INTERNALISM*) For all *r*(*F*,*P*,*A*,*C*), *r(F,P,A,C)* is a reason relation holding between a fact *F* and a paradigmatic agent *P*'s action *A* in circumstances *C* iff (and because) *r*(*F*,*P*,*A*,*C*) holds in virtue of the desires that feature in *P*'s idealized psychology.[2]

Some comments are needed. When I mention facts *F*, I am indifferent between talking about (true) propositions, instantiated properties, or whatever else one may take facts to be.[3] By a 'paradigmatic agent' *P*, I mean someone who has beliefs, desires, and is instrumentally rational, as per *PARADIGMATIC AGENCY*$_{IR}$ in chapter 5. This is because I am primarily interested in such agents – but I shall relax this assumption when discussing the shmagency objection in chapter 7, section 4. Moreover, what an action *A* is should be unproblematic by now, and I discuss the circumstances *C* later (cf. chapter 7, section 2).

But what does 'the desires that feature in *P*'s idealized psychology' mean? Even though reasons are explained by *P*'s desires here, Humeans tend to agree that some level of idealization of *P*'s psychology is necessary to get our reasons right. Perhaps, for example, some of our desires are short-sighted and incompatible with some of our other,

[2] *r*(*F*,*P*,*A*,*C*) may hold between more *relata*, e.g. times, as well. But I will stick with this formulation for simplicity.

[3] Since I will argue that reasons can feature in reasoning, one might think that propositions would be the best idea here. But I have no axe to grind in the debate, so I shall be open to more metaphysically heavy-duty views if they can do the relevant work (cf. Peter, 2019).

deeper, desires, or our desires are based on faulty information. In better conditions, our desires might explain our reasons better.

From here and on, I shall follow convention and call fully idealized agents '*A*+', while non-idealized ones will be called '*A*'. As paradigmatic agency involves beliefs, desires, and instrumental rationality, and as desires do the key explanatory work here, the core idea behind *DESIRES INTERNALISM* is that if *A*+ desires to φ, then *A* has a reason to φ. And the idealization of *A*'s psychology, i.e. that which turns *A* into *A*+, plays a supporting role in ensuring that *A*+ has the *right* desires to explain why *A* has the reasons for action that she does.

How does that work? As *DESIRES INTERNALISM* explains reasons by appealing to the desires that feature in an idealized psychology, and that psychology is functionalistic, I shall take an ideal agent to be a fully functional paradigmatic agent, in the sense that all the psychological states and capacities she has *qua* paradigmatic agent function fully, and that she manifests these capacities fully insofar as she acts. I shall also assume that, *qua* ideal agent, she has or is in the right background conditions to be able to exercise those well-functioning psychological states and capacities. Here, a 'background condition' is any fact about the agent or the context she is in which may affect her psychology, decisions, or actions.

To be idealized, then, *A*+ is supposed to feature all the properties of *A* just mentioned, fully functioning or fully manifested when acting. So, for example, the idealized agent does not just have the capacities for believing, desiring, or instrumental rationality, but actually is instrumentally rational insofar as she acts (cf. Smith, 2012*c*). This idealization condition rules out the possibility that the mental states involved in idealization are blocked so the agent is unable to make use of them, e.g. by accidie (cf. Hurtig, 2006).

Second, full idealization also requires that *A*+ has or is in the background conditions that allow her to have the states or capacities of her psychology fully functioning or manifested when acting. The point here is that the agent cannot be in background circumstances that do not allow her psychology to fulfil its functions. For example, insofar as having capacities to deliberate requires ways of applying the instrumental principle in deliberation, and deliberation is something that an ideal agent may do, having the abilities to apply the instrumental principle in those ways is an idealizing condition of the agent. As Williams (1981) famously noted, instrumental rationality need not just involve the taking of already known means to given ends, but can also involve deliberation about how to do so by finding constitutive solutions in cases where desires conflict, using one's imagination to find new possible solutions, etc.

This kind of background conditions idealization can also explain why *A*+ ought to be epistemologically refined. Epistemic refinement does not just mean that *A*+'s beliefs are fully functional or manifested so that she satisfies the epistemic aim of belief (assuming that that is truth, knowledge, justification, or something else). Just by assuming the full functioning of beliefs, I shall assume that, insofar as the ideal agent has beliefs, these live up to their aim.[4] However, many authors have also tended to assume that the agent must have some particular set of beliefs, e.g. all relevant true beliefs and no false ones (Smith, 1994, ch. 5). I am not sure exactly what set of beliefs is necessary for ideal agency, but it is plausible that there is a set like that, for it is plausible that *A*+ needs fairly many true beliefs to manifest her agential capacities fully insofar as she acts – without them, she might take the wrong means to her ends. And if there is such a set, *A*+ has it. Summing up, then, idealization involves fully functioning or manifested capacities *as well as* relevant background conditions. The latter is also what explains (*inter alia*) why *A*+ is able to apply the instrumental principle in complex ways and is epistemically refined.

The next important aspect of *DESIRES INTERNALISM* is that it says that reasons hold *in virtue of* the desires that feature in agents' idealized psychologies.[5] When I introduced *FORMAL CONSTITUTIVISM* back in chapter 1, section 2, I briefly mentioned that there is some controversy about what it would mean for something to explain (or back) something metaphysically, but also that the question of which relation obtains here is orthogonal to the main constitutivist project. What matters for constitutivist purposes is whether a normative phenomenon holds and has normative force for some agent in virtue of some properties of some constitutive feature(s) of some aspect of agency. But since idealized agency is constituted by idealized beliefs, desires, and instrumental rationality, *any* explanatory (or 'in virtue of'-)relation that holds between them and some reason can do the relevant job.

This point can be put using the abbreviations from chapter 1, section 2: if some reasons *G* are reasons in virtue of the properties of some feature constitutive of an

[4] Smith (2012*c*) has a similar argument for taking the beliefs of the ideal agent to be true. Because beliefs aim at truth, perfectly held beliefs will be true.

[5] For the condition, see Finlay and Schroeder (2017; cf. Schroeder, 2007). Hurley (2001) points out that reasons internalism is compatible with some sort of Platonist robust realism about the good or reasons, since recognizing the good would involve forming pro-attitudes for it. But Hurley's discussion is not formulated in terms of an explanation; if we only read *DESIRES INTERNALISM* biconditionally, her point holds, but if we read it as 'iff (and because)', the distinction between internalism and externalism becomes much clearer.

idealized agent *F*, and the relation between *F* and *G* holds between them, the reasons (*G*) can be said to be explained by the desires of the idealized agent (*F*). But several different explications of the explanatory relation are possible here. It may be that reasons are grounded in (or 'sourced in', as Chang (2009; 2013) might say) the desires of an idealized psychology. Or they may be constituted or realized by them, or even identical to them. Constitutivism itself does not say.

For now, however, I shall try to formulate my framework in as orthodox a manner as possible. This means that, in this chapter and the next, I shall say that reasons relations have their *sources* in the desires that feature in the idealized psychology of *A*'s ideal counterpart *A*+, or, equivalently, that reasons are *grounded* in these desires. Grounding relations are typically taken to be asymmetric (so just because some property *G* is grounded in some property *F*, *F* is not grounded in *G*) and hyperintensional (so even though *F* and *G* may be necessarily co-extensive, they need not be identical), allowing that the grounded property *F* is something over and above *G*. Hence, for present purposes, I shall interpret the reason relation *r*(*F*,*P*,*A*,*C*) as holding because it is grounded or sourced in the idealized features of *A*+'s psychology.

So much for the two most important explanatory properties. It is the idealized desires that feature in the psychologies of ideal agents that explain their actual counterparts' reasons, and the desires do so because they are the sources or grounds of reasons. This type of theory fits the constitutivist schema from chapter 1, section 4. *DESIRES INTERNALISM* is a version of:

> (*FORMAL CONSTITUTIVISM$_P$*) An account *T* of some normative phenomenon P_P is constitutivist iff *T* entails that P_P is normative because P_P is, or is normative in virtue of, some property or properties of the constitutive feature or features *C* of an aspect of agency *A**, where *C* constitutes something as an *A**.

According to *DESIRES INTERNALISM* (*T*), idealized desires explain what it is for a practical normative phenomenon – normative reasons (P_P) – to obtain. This is because reasons are grounded in the constitutive features (*C*) beliefs, desires, and instrumental rationality of *A**, where that aspect of agency is (idealized) agency.

(2) Explanations

As mentioned, the positive part of my argument will be explanatory. I will argue that the properties of an idealized psychology that feature in *DESIRES INTERNALISM* can capture a whole lot of features of reasons that we want a theory of normative reasons to capture. This means that it is very plausible that practical reasons indeed are explained by idealized desires *qua* constitutive features of idealized agency. The main case I shall make for that conclusion is cumulative; there are many features of reasons that a theory of reasons should make sense of, and no features are fundamentally privileged over others – what matters is that a large amount of them can be explained.[6]

So which features of reasons for actions should be explained? And why some rather than others? I take the features that matter to be a collection of commonly accepted ideas about reasons in the literature. I see these features as 'data' for theories of reasons to explain. In fact, sometimes, the features that I set up to be explained are brought up as counterarguments to Humean theories (on the assumption that Humeans cannot explain them). But I shall show how *DESIRES INTERNALISM* indeed can do so.

I will not, however, discuss features of reasons that Humeans (or others) can explain almost trivially. Examples of these are easy to come by: humans have reasons (because, Humeans think, humans have desires), we can be mistaken about reasons (because we can be mistaken about our desires, not to mention idealized ones), reasons supervene on other facts (because they co-vary with them over worlds), etc. A theory is better if it explains *interesting* features of the phenomenon it is supposed to explain.

Nor will I try to explain features of reasons that Humeans can capture easily but where several theoretical options are open for them. The features I have in mind here include, for example, how reasons vary in strength, how some reasons may be optional

[6] This general explanatory line of argument is one of several typically adduced to defend Humean theories of reasons (cf. Schroeder, 2007). There are, however, many other arguments in the literature (cf. Brunero, 2017; Heathwood, 2011, for overviews): (*i*) reasons must be able to *explain* actions (Williams, 1981), (*ii*) a reason must be able to *motivate* actions (Korsgaard, 1986; Markovits, 2014, pp. 65-66), (*iii*) some form of reasons internalism is a conceptual truth as an analysis of the concept 'reasons' (Williams, 1981), (*iv*) this explanation is analogous to arguments for theoretical reasons (Goldman, 2011; Markovits, 2014, pp. 58-65; cf. Velleman, 1996), (*v*) internalism makes good sense of our (social) practices of reasoning (Manne, 2014; Strandberg, 2018), (*vi*) it follows from some understanding of deliberation (Paakkunainen, 2018), and (*vii*) it explains failing to act for one's reasons well (Markovits, ch. 2014, pp. 54-58).

Even though my case for *DESIRES INTERNALISM* is cumulative, versions of (*i*), (*ii*), (*vi*), and (*vii*) are at least some of the features I adduce in favour of internalism. I discuss (*i*), (*ii*), and (*vii*) when I discuss *PRACTICAL SIGNIFICANCE* below, and (*vi*) when I discuss *FIRST-PERSONALITY*. However, I doubt (*iii*), i.e. that *DESIRES INTERNALISM* is a conceptual truth – it seems far too contested for that. And I am silent about the explanation of theoretical reasons (*iv*), as well as about what our practices of reasoning are (*v*), though I suspect the latter can be in part explained by the theory defend.

while some are not, or how instrumental reasons should be understood (cf. section 1). As far as these features go, the theoretical preferences any particular Humean has have no impact on the general question of whether reasons have their sources in idealized desires. Instead, I hope *DESIRES INTERNALISM* will be shown to be powerful because it explains features that have bearing on what reasons have their sources in, not features Humeans reasonably may disagree on.

With all that said, let us now get started with the explanations.[7]

> (*RELATIONAL CHARACTER*) Reasons are reasons *for* an agent, in the sense that they and their normative force are closely related to the agent and her motivations.

Explaining *RELATIONAL CHARACTER* involves explaining how reasons matter for agents in a non-alienated manner (cf. Railton, 1984). There is significant disagreement about how the explanation of the relation between the normativity of reasons and agents' motivations should go, however (cf. e.g. Dreier, 2015; Korsgaard, 1996*a*, ch. 1; Paakkunainen, 2017; Parfit, 2011*b*, pp. 413-425).

Still, to put the point I think lies behind *RELATIONAL CHARACTER* in general, the idea is that for an agent not to be alienated from her reasons, her reasons must matter for her in a way where they and their normative force are, in *some* way, closely related to her and what she can be motivated to do.

On *DESIRES INTERNALISM*, the explanation of *RELATIONAL CHARACTER* stems from how reasons depend on *A*+'s psychology. Most significantly, reasons have their sources in agents' idealized desires, where the idealization shows how they have normative force for *A* (cf. *PRESCRIPTIVITY* below). As the reasons depend on *A*'s idealized desires, they are closely related to her.

Moreover, internal reasons also tend to be motivationally close to agents' motivations. Many of *A*'s present desires are likely to remain after idealization, and she is also likely to form new desires based on her appreciation of her reasons over time (cf. *PRACTICAL SIGNIFICANCE* below). Hence, *A*+'s desires are closely connected to *A*'s both in general and motivationally – *A*+'s desires are simply those of *A*, plus.

[7] Note that the first few properties on my list come from Scanlon (2014, ch. 1). I diverge from him by not discussing whether reasons have non-subjective truth values or supervene on natural properties, however – those are some of the issues that seem to stem from orthogonal debates.

(*KNOWLEDGE*) We often know the reasons we have.

It is widely recognized that we often know what our reasons are. Some philosophers go further here, however: for example, Wiland (2012, p. 131) thinks that theories of reasons 'should be able to account for how an agent ordinarily knows 1) her reason, 2) the action, and 3) the connection between the two, all 4) immediately, and 5) spontaneously.'

However, I prefer my weaker, and fairly vague, formulation of *KNOWLEDGE*. Wiland's conception of our knowledge of our reasons seems too strong. The extent to which we know the reasons we have plausibly varies both with our epistemic abilities and with our circumstances. It is possible that our ability to come to know our reasons may be deficient, or that we are in circumstances that we fail to know well, so it is hard to say to what extent we know our reasons in general.

Even so, there is a puzzle here. For assume that our reasons indeed are grounded in our idealized desires. Even though our idealized desires plausibly usually are fairly similar to our present desires, they need not be the same. So insofar as our reasons are supposed to be explained by our idealized desires, it is not obvious how we can know what they are.

I think the proper solution to this puzzle involves several different elements. Some reasons are individual, some are universal, and our knowledge of them differs. Individual reasons depend on highly variable idealized desire sets, so they are often or even usually different for different agents. Universal reasons, however, depend on universally required desires (cf. chapter 7, section 3).

When it comes to individual reasons, we can make *some* good on the suggestion that we know our present desires, or what we may desire if we were to reflect on them, and hence our reasons.[8] Sometimes we already have the desires that would survive after due idealization. But sometimes we do not: some desires can be lost, and others gained, over the course of an idealization process. We need to see how agents can have evidence of what desires the ideal agents would have after due idealization, and then see how agents can appreciate that evidence to acquire knowledge (of their reasons).

The evidence comes from the fact that the desires of idealized versions of us *qua* agents ground our reasons. Because they do that, to the extent that we approximate being

[8] This gives desires an epistemic role. This is not an uncommon point to make, even for Humeans (cf. Schroeder, 2007, pp. 155-163). Does this mean that I assume that desires aim to represent the world accurately? No, their function is to explain our actions (and they also explain our reasons), but it is not part of their function to give us knowledge. Rather, their epistemic role *piggybacks* on their functional and explanatory roles. We can use them as a heuristic (cf. appendix A, section 1).

ideal, we also approximate having the desires that ground our reasons. In the limit, the desire sets will be co-extensive, for the desires that we will have at full idealization will be the same as those that hypothetical, idealized versions of us have. And there are many types of desires that plausibly survive idealization. For example, if I keep desiring to watch Tottenham Hotspur play their next game, year in and year out, through many different life circumstances, then presumably, I will retain a desire to watch them play their next game at full idealization. If so, knowing that I have a desire to see them play their next game is enough for me to know my reason.

Moreover, many desires are so general or abstract that, barring exceptional circumstances, we will not lose them. These include desires for air, food, and survival, and probably the others on some list of Rawlsian primary goods as well (cf. Rawls, 1971, pp. 60-65; 90-95). Finally, we can usually infer what people generally will desire based on facts we know about the life situations they are in. For example, it is well-known that increases in material welfare seems to generate more expensive tastes, as people adjust their tastes to their earnings. It is plausible that we can know our reasons in various situations based on such general knowledge, too.

But we can still fail to know what our reasons are. We need not, at present, have perfect insight into the desires that fully idealized versions of us *qua* agents might have. And to see how we can acquire evidence of what the desires or reasons we would have after idealization are, we need to say more than that we know our present desires. We need to see how we can gain reliable knowledge of them even though our ideal desires may differ from our present ones. To that end, my proposed explanation is a feedback process originally proposed by Railton (1986).

Railton defends a model of a person's good where that is what their epistemically and instrumentally idealized counterparts would advise their non-idealized counterparts to do if they were in their non-idealized counterparts' circumstances and took their present desires into account. This model is obviously similar to *DESIRES INTERNALISM*. Railton thinks that the process of attaining knowledge of one's good need not be 'one of an ideally rational response to the receipt of ideal information, but rather of largely unreflective experimentation, accompanied by positive and negative associations and reinforcements' (Railton, 1986, p. 180). The idea is that insofar as there is a standard of the good – or our reasons – which is independent of what we think, and the world gives us feedback about how we attain it, our desires and knowledge of them will tend to start to approximate the good – or our reasons. This does not rule out engaging in more theoretically complex deliberation and desire-acquisition, but it is not necessary to do so to acquire (reliable) knowledge about what our ideal desires (and hence our good, or our reasons) are. The feedback process can help regardless.

Thirst-based reasons are a good example of the process in action. I feel my throat is parched, so I lose my concentration when writing, and start to think about what to drink. Plausibly, that says something about the desire I would have given sufficient idealization: I would know that I need water not to become dehydrated, and in the long run even to survive. As I already want to survive, I develop a desire to drink water. In this way, my situation gives me feedback about what my reason is.

In fact, because *A*+'s desire set is an idealized version of mine, it is likely that the more feedback I acquire about my situation, my own desires will also approximate *A*+'s desires. This is because *A*+'s desires are my idealized desires, not just any old external desire set. So the connection between our respective desire sets makes the feedback-style explanation more plausible – the feedback process may *cause* the idealized desire set that my beliefs about my reasons should correspond to.

It turns out, then, that going through an idealization process, and hence approximating *A*+, will get us closer to our reasons, even though that rationalization process need not be an intellectual or theoretical endeavour. Sometimes we will retain our older desires (or can infer which ones we are likely to acquire), and sometimes we acquire new ones through feedback processes. So there are several sources of evidence about our individual reasons.

However, universal reasons differ from those explained by the desires mentioned here. There is no reason to believe that procedurally rationalizing desires will give everyone the same desires, and hence reasons, in their different individual circumstances (*pace* Smith, 1994, ch. 5; cf. Leffler, 2014, ch. 1). Universal reasons have a different epistemology, for the connection between the desires that ground them and our own desires when fully procedurally rational has another basis. In chapter 7, I shall argue that ideal agents constitutively possess a desire to cooperate with other cooperating agents. The desire to cooperate explains universal reasons (together with the other desires they cooperate to satisfy). If that is right, we can get to know what our universal reasons are, and how they impact the structure of our other reasons, by transcendental reflection.

The idea, then, is that we can get to know what our universal reasons are by reflecting on which desires are rationally required, not by rationalizing our current desires. But that reflection is much more abstract and theoretical than the knowledge we have of the desires that ground individual reasons.[9] This means that the epistemology of reasons has two fundamentally different sides. We are often able to know our individual reasons by consulting our own desires or feedback processes, and our universal reasons through more abstract arguments about which desires are rationally required.

> (*PRACTICAL SIGNIFICANCE*) Agents can be, and often are, motivated to act for their reasons.

This feature of reasons has, famously, often been formulated in terms of an internalist motivational constraint. According to the constraint, if an agent cannot be motivated by a consideration (at least after due idealization), the consideration does not count as a reason for that agent (Williams, 1981). But this formulation of the constraint is also famously controversial (Bedke, 2010; Markovits, 2014, ch. 2; Millgram, 1996).

DESIRES INTERNALISM is compatible with that constraint, but to avoid getting stuck on Williams' argument, I formulate *PRACTICAL SIGNIFICANCE* relatively weakly by saying that reasons *often* motivate actions. The present formulation is, I think, common ground between most writers in the debate, whether internalist or externalist – and my explanation of motivation from reasons will be as well.[10] It does not, however, matter for my purposes that reasons externalists *can* accept the conclusion here; what matters is that I can show that there is a plausible way in which motivation from (normative) reasons can make sense on a Humean picture.

One aspect of this connection comes from the relation between the desires that ground normative reasons and the desires that feature in motivating reasons. Humeans famously distinguish between motivating reasons, construed as belief/desire-pairs, and normative reasons (cf. chapter 4, section 1; appendix A, section 3). The idealized desires that explain normative reasons according to *DESIRES INTERNALISM* may, of course, be extensionally equivalent to the desires that feature in the belief/desire-pairs that make up motivating reasons. So the idealized desires may be equivalent to actual desires, and can hence feature in action explanations.

[9] In fact, since I shall argue that we have universal reason to cooperate to ensure the satisfaction of our other respective desires, it will also turn out that *others'* desires or reasons will matter for us. But I take it to be uncontroversial that knowledge of others' desires can be added to our knowledge of our own.

[10] One externalist who agrees with me is Enoch (2011*a*, p. 225). He thinks that *A* φ:s for reason *F* just in case: (*i*) *A* intentionally φ:s, (*ii*) *A* believes that *F*, and (*iii*) the belief that *F* is a (normative) reason for *A* to φ in the circumstances plays an appropriate causal role in bringing about *A*'s φ-ing. This is *exactly* my story.

But *how* can they come to be actual desires? Assuming that the case where we actually have them and they are (correctly) linked up with beliefs to form instrumental desires is unproblematic, there are still many cases where the desires and beliefs are not linked up, or we do not actually have the desire sets that we would have after suitable idealization.

Moreover, there also seems to be a further constraint that a plausible explanation of acting for normative reasons must meet: we must be able to say how agents act *for* their reasons, not just in line with them. It is not enough just to have a belief/desire-pair that would have one do what one has reason to do. Not least, such belief/desire-pairs could be causally deviant. Hence, some philosophers have argued that the relevant belief/desire-pair must be put together in virtue of the agent's recognizing her normative reason (cf. e.g. Korsgaard, 2008, p. 63). And in the light of *inter alia* Korsgaard's considerations, Arpaly and Schroeder ask: 'How can thinking and acting for reasons be a causal process, when causation by attitudes with rationalizing contents does not guarantee thinking or acting for reasons?' (Arpaly and Schroeder, 2012, p. 61)

To explain how the desires that explain normative reasons can feature in motivating reasons, I largely follow Smith (1994, ch. 5).[11] Consider an agent who is deliberating about what to do. She weighs up various reasons, compares and contrasts them, and finally settles on a course of action. I shall present two causal hypotheses about how her recognition of what she takes to be her normative reasons can explain the motivating reasons behind her actions.[12]

First things first, however. What is it to *recognize* a reason? To recognize that one has a reason to φ – or, in other words, that a reason relation $r(F,P,A,C)$ obtains – is to believe that F supports φ-ing when one's cognitive capacities are suitably well-functioning. In the limit, being suitably well-functioning means, at least, that one's mental states work according to their functions, the relevant background conditions are in place, and one is suitably epistemologically refined. (Hence, these factors overlap with the idealizing

[11] However, Smith's view is slightly more theoretically complex than mine. He considers intentional explanations of actions (in terms of motivating reasons) to differ from deliberative explanations (in terms of normative reasons). But I do not want to take a stand on whether these explanations are fundamentally different; given the connections between normative and motivating reasons that I defend here, perhaps they are both part of a larger overarching explanation.

[12] Enoch (2011*a*, pp. 234-235) mentions several others. One possibility is that normative reasons link up with a (possibly fetishistic) *de dicto* desire to do what we have reason to do (though cf. footnote 14 below). Another is that we sometimes have other desires which are such that responding to reasons is a good policy for satisfying them (though this would make acting for reasons strangely instrumental).

Other possibilities still involve even more downstream changes; a belief may bring about changes in the strengths of desires, or the acquisition of a belief might change other beliefs that already might feature in an agent's belief/desire-pairs. I presume that these options are not the main ones.

conditions on ideal agency from section 1.) But we need not be in the limit to be suitably well-functioning; it is enough that we approximate the limit to be able to recognize a reason.[13] Moreover, importantly, one need not believe one has a reason under that description to recognize it; perhaps one only recognizes a consideration that makes an outcome appealing. As a matter of phenomenological fact, that often seems enough for motivation.

With recognition described, we can introduce the hypotheses. The first one emerges from the case where the agent already has the desires that explain the reasons she recognizes. If so, her recognition of her reasons plausibly causes her capacity for instrumental rationality to combine those desire with some means-ends belief(s). When I defended IR_{HUMEAN} back in chapter 5, section 3, I did not say anything about what would explain why or when that capacity would cause instrumental desires, but it seems extremely plausible to think that there is some causal link between our judgements about what we have reason to do and becoming motivated to do that.[14] It is a fact of life that we tend to become motivated by what we think we have most reason to do.

Moreover, it is often the case that we already have the reason-explaining desires, as per the discussion of *KNOWLEDGE*. The idealization process that I have assumed is fairly minor; idealized agents will often retain many of their original desires at the end of their idealization processes. So it is often the case that one already is motivated to act for one's reasons.

But then we get to the next, trickier, case. If the agent does not already have the desire that she would have after idealization, how could she form it? Here my causal hypothesis is that our recognition of our reasons can make us form desires as well as trigger applications of our capacity for instrumental desires. Recognizing that we have a reason to φ, in turn explained by our hypothetical motivation to φ, will sometimes *cause* our actual motivation to φ.

Is this plausible? Consider Railton's feedback mechanism again (cf. *KNOWLEDGE*). It seems likely that insofar as our reasons are settled by ideal desires, the world provides feedback that may have impact on those desires. Williams (1981) famously thought that, despite being thirsty, one would not retain a desire to drink a glass

[13] So why the idealization, then? It seems implausible that *merely* believing there is a reason will do any work. An epistemically refined agent is likely to be more in touch with the facts of the matter: more reliable, more accurate, more responsive, etc. So her belief is likely to penetrate her psychology further.

[14] To be clear, while this link sometimes may be based on a *de dicto* desire to do what is right, which Smith famously calls fetishistic (Smith, 1994, pp. 71-76), it need not be. It is just as possible that we simply have dispositions or capacities to form desires in the light of our judgements.

of clear liquid in front of one if one knew that the glass contains petrol instead of – as one might have assumed – gin. But assume that, instead of gaining knowledge that the glass contains petrol prior to drinking it, one actually tastes the liquid. Recognizing that it does not taste like gin, one immediately forms a desire not to drink the liquid. That desire is presumably equivalent to the desire one would have after idealization; we have reason not to drink petrol because we are averse to it because it would be bad for us.

We can interpret the aversive desire generated by tasting the petrol as an instance of Railton's feedback mechanism. The taste of petrol makes one recognize the reason not to drink the liquid, which in turn generates a desire not to drink it. The original reason is a reason in virtue of what one's desire would be had one been suitably informed, but now that the agent in the case becomes suitably informed by tasting the petrol, she forms a desire which fits the facts about her reasons. The new desire may then link up with suitable beliefs, ensuring that the agent does not drink the petrol.

This feedback mechanism-style explanation does, though speculative, seem fairly causally plausible. Moreover, there is no reason to believe that our ability to generate desires in virtue of idealization would have to be this concrete. It can be generalized; it may well be that we can form desires in virtue of gaining more knowledge by more abstract means, or by reasoning about how to satisfy our other desires, as well. This point, too, fits the phenomenological facts. If I come to believe that φ-ing is wrong, I tend to not want to φ.

If the two causal hypotheses here are correct, it follows that the desires that explain our normative reasons also can feature in our motivating reasons. Moreover, they can do so because of our recognition of normative reasons; if we recognize the reason, we can *ipso facto* form the relevant desires (and have them link up with appropriate means-beliefs). Hence, to summarize, even when we do not have the right belief/desire-pairs prior to engaging in reasoning, desire-based normative reasons can come to feature in actions as motivating reasons in two ways when we recognize them. First, they can trigger the causal mechanism that unifies pre-existing desires and beliefs. Second, they can cause desires, which then link up with appropriate means-beliefs.[15]

So much for successful cases of action for reasons. But what happens when recognizing a reason motivates an agent improperly, or motivates her via a deviant causal

[15] One may wonder whether this explanation commits me to some kind of motivational internalism about reasons judgements, where these should be construed as motivating. I am not sure. I do not want to say that reasons judgements *always* are motivating, but it *may* be possible to interpret my view as some sort of conditional internalism. At present, however, I am not committed to that.

chain? Such cases do, in fact, speak in favour of my explanatory hypotheses. It is plausible that we form desires, final or instrumental, based on what we *take to be* reasons, which may or may not correspond to what actually is there. We may act for no normative reason whatsoever, or for very poor ones. If that is right, it may well be the case that we become motivated to act for no, or at least poor, reasons. Such cases are common in real life, and their existence therefore a fact to be explained. Moreover, the desire-generating ability of reasons recognition may malfunction, and – for example – generate too weak desires, or possibly no desires at all, for us to be motivated to do what we have reason to do. That seems like a very good explanation of at least some cases of *akrasia* – and *akrasia*, too, is a phenomenon to explain.[16]

Second, what about cases of deviant causation? Well, plug in your preferred solution to the problem. My framework allows for many different possible solutions (cf. appendix A, section 2). It is possible to slot in pretty much whichever explanation one favours here and have it come out right. It will just be an add-on to how one interprets the causal process taking place between the recognition of a reason and the formation of an instrumental desire.

Finally, I want to indicate how the causal hypotheses I have suggested explain one further phenomenon. It seems puzzling that normative reasons – which are factive – may motivate actions on Humean views, as motivating reasons here are belief/desire-pairs, but normative reasons are not psychological. But the causal hypotheses here say that it is the agent's *recognition* of her reason that triggers her instrumental rationality or her desires plus instrumental rationality. And recognizing a reason is something we do with our mental states. So the explanation seems to get the relation between factive normative reasons and psychological motivating reasons right.

> (*FIRST-PERSONALITY*) A theory of reasons should explain how reasons phenomenologically appear to feature in first-personal deliberation. This involves, at least, the claims that: (*i*) reasons feature in deliberation, and (*ii*) we reason using their contents, rather than using mental states such as desires.

Both (*i*) and (*ii*) seem very plausible. It is often taken for granted that deliberation proceeds using reasons, or at least that when we deliberate well, we do so with reasons.

[16] It is even possible that we should understand what I called execution-*akrasia* (in chapter 5, footnote 15) in terms of a failed reasons recognition. That would complete my explanation of *akrasia*.

Among other things, deliberation involves weighing reasons for and against different courses of action (Kolnai, 1961). So as per (*i*), reasons feature in deliberation.

Similarly, it is often assumed that we deliberate using the *contents* of our mental states rather than the states themselves, even though reasons feature in our deliberation (cf. Schroeder, 2011). It seems to be what we have reasons to do, or *F* in *r*(*F*,*P*,*A*,*C*), that features in deliberation, not the desires that ground reasons. But how could that be the case when reasons are explained by desires?

An obvious move to make regarding (*i*) is to say that desires need not feature in our phenomenology – indeed, one property of desires might well be that they make the *objects* one desires look attractive (cf. appendix A). Instead, their work is *backgrounded*, so from a first-personal perspective, one need not – though one maybe sometimes can – have them phenomenologically occurent in one's deliberation, whereas one actually deliberates using representations of facts (cf. Pettit and Smith, 1990; Smith, 1994, ch. 4). This is plausibly what happens insofar as desires are the grounds of our reasons. They settle which reasons there are, without therefore themselves necessarily *being* the reasons.

This also feeds into (*ii*), i.e. the fact that reasons are factual and have their reasonhood explained by being what one would desire upon idealization, rather than actually desires. It is easy enough to let facts – understood as, for example, true propositions – feature in one's deliberation, without therefore saying that desires do so.

> (*PRESCRIPTIVITY*) Reasons for action are prescriptive, in the sense that they have normative force. Hence, they prescribe actions by, for example, justifying or requiring them.

Again, by 'prescriptivity', I mean most generally that reasons have some form of prescriptive force or normative oomph (cf. chapter 1, section 2; chapter 5, section 5). Many philosophers have emphasized that reasons have this property (e.g. Parfit, 2011*a*, pt. 1; Star, 2016, ch. 2; cf. Persson, 2013, ch. 12).

But while I have defended directivity as an account of the normative force of instrumental rationality, that form of normativity does not need to be relevant for reasons. Most of the properties of the normativity of rationality carry over to the normativity of reasons – they are different from mere feelings or wants, they appear constraining for choice, they ground criticizability, and we can act for them or fail to act for them.

However, reasons aggregate with each other, and we cannot respond to our reasons if we fail to recognize them in the way that we might be rational even if we fail to respond to our reasons, so the normativity of rationality is not quite the same as that of reasons. Moreover, it is possible to act even paradigmatically without acting for any normative reason whatsoever (cf. Stocker, 1979). Hence, they are not plight inescapable, so an explanation of their normativity along those lines would not work.

Instead, the *PRESCRIPTIVITY* of reasons stems from the fact that they are grounded in rationalized desires. Rationality is a feature of the idealization process, and their normativity stems from rationality. The facts about these desires are *legitimized* as normative by having been accepted by due rational process (cf. Korsgaard, 1996*a*, ch. 3). Legitimized desires have a special status. As instrumental rationality is normative, the desires that one would have if one would be fully instrumentally rational are such that they, too, have a special status as legitimized.[17]

An analogy might be helpful to bring out this point. Plausibly, a procedurally just law is procedurally just iff (and because) it has been formed by due political process, such as having been passed by a parliament elected in fair democratic elections. Similarly, a desire that one has iff and because one is duly rational becomes legitimized because one's psychology is legitimately constituted in virtue of conforming to the structural normative demands that constitutively hold for an agent. Metaphorically, rational desires are the agent's executive powers to make facts reasons, her pre-rationalized desires are the legislative assembly, and the rationalization procedure serves as her judiciary.[18]

Hence, the normativity of reasons has been explained by something else that is normative (i.e. instrumental rationality). This differs from the kind of normativity possessed by instrumental rationality – instrumental rationality is plight inescapable. However, legitimization can still explain the relevant properties of normativity. Legitimized desires differ from ordinary desires, feelings, or wants just because they are

[17] This legitimization procedure does not just positively explain prescriptivity. It also helps solving some other, fairly minor, problems for Humean theories of reasons. According to Schroeder (2007, ch. 3; 10), two problems for Humeans are: (*i*) that a Humean theory risks being either *inconsistent* – one needs a reason to act on desires, but all reasons are explained by desires – or *chauvinist*, because it presumes just one non-desire-dependent reason, which is to take the means to one's ends, and (*ii*) that it is objectionably instrumentalist, because it takes desires to be, in some sense, given or not subject to rational criticism.

Both these problems dissolve if reasons stem from desires that are legitimized. (*i*) disappears because one need not have a *reason* to act on desires. The normativity of taking means to ends is explained by the normativity of instrumental rationality, which does not involve reasons for action. This explanation is neither inconsistent nor chauvinist; reasons are another kind of normative entity than the principle of instrumental rationality. Moreover, (*ii*) disappears as the legitimization procedure is rich enough to allow that no desire is given. They must all be accepted by one's rational idealization. The Humean need *never* take existent ends for granted.

[18] This view can be contrasted with Korsgaard's constitutional model of the self, where the self also stands behind actions (Korsgaard, 2009, ch. 6-7; cf. chapter 2, footnote 7). The self does a similar procedural job for her when it legitimizes maxims by *CI*, but here, we legitimize certain possible desires by IR_{HUMEAN}.

legitimized. They are constraining for choice because insofar as we deliberate, we do not merely take our desires, wants, or feelings to matter, but rather something with a special status, such as that of legitimacy. They show why someone who fails to act on them is criticizable: an agent who does not act on her reasons fails to act on considerations grounded in her legitimized desires, and hence acts from desires that are not legitimized. That seems less than ideal. And, of course, we can either act, or fail to act, for our reasons. So all the features of normativity that reasons plausibly have are captured here.

> (*UNIVERSAL/INDIVIDUAL*) Some reasons justify actions for separate individuals, whereas some justify actions universally.

The formulation of *UNIVERSAL/INDIVIDUAL* straightforwardly captures what it means. This property has been emphasized by a number of recent internalists (e.g. Schroeder, 2007; Strandberg, 2018), though it also has been famously denied by many older ones in the literature (e.g. Williams, 1981). Intuitively speaking, however, it seems fair enough to assume that it pre-theoretically is the case that some reasons are shared by all agents. Moral reasons seem to be shared by all, but reasons to drink beer or wine do not seem to be.

A straightforward explanation of *UNIVERSAL/INDIVIDUAL* can be given with *DESIRES INTERNALISM*. Desires that are rationally required (or at least rationally required under standard human conditions) are the grounds of universal reasons, whereas those that are not so required are the grounds of individual reasons. This is because all fully rational counterparts of agents have the former desires, but different desires among different agents explain the latter. In line with this explanation, the reasons discussed in this chapter are individual, but I shall go on to explain some universal reasons in chapter 7.

> (*EXTENSION*) One cannot have reasons to do just about anything, nor should we assume that we have no reason to do what we usually think that we obviously have reason to do.

The terminology here comes from Schroeder (2007, ch. 5-6), but objections like *EXTENSION* have been raised against subjectivist theories about most normative phenomena throughout the history of philosophy. In the context of *DESIRES INTERNALISM*, many have argued that internalism has the wrong normative implications because it may provide *too many* reasons, such as reasons to count blades of grass (Rawls, 1971, p. 434) or reasons to cause ourselves future agony (Parfit, 2011, pp. 73-82) if those are things we desire. Similarly, *DESIRES INTERNALISM* may provide *too few* reasons, such no reasons to be moral for Sensible Knaves, Fooles, or Gyges, or no reasons to avoid future agony if we are indifferent to that.

We can construe these objections as challenges to the extensional adequacy of theories like *DESIRES INTERNALISM.* Admittedly, the extensional fit between the reasons explained by a general theory of reasons and our pre-theoretical intuitions need not be too strong, for we should not presume that we, pre-theoretically, know *exactly* which reasons we have. That question is for normative inquiry to settle, and our assumptions of normative knowledge often appear suspicious (cf. chapter 1, section 3). However, we should not be complete sceptics about some intuitions about what we have reason to do either. General worries about philosophical methodology aside, the epistemic challenges I launched there are primarily challenges for the moral knowledge we presume we have, not for our knowledge of our individual reasons. So getting at least fairly intuitively plausible results seems important for a theory of reasons.

The issue, then, is not fundamentally whether there are too many or too few reasons per se. We may have very many, or very few, desires. The problem is rather that our desires may be normatively problematic in the ways they seem to be in the classic examples two paragraphs ago.

For now, I shall bracket morality. I will return to it in chapter 7. Morality aside, however, I shall argue that there is no reason to think that a pre-theoretically odd desire set is normatively problematic in the way that is assumed in the *EXTENSION* objection. My argument for that conclusion is inspired by Derk Pereboom's (2001, ch. 4) Four Cases argument against compatibilism about free will. My aim will be to show that desire shifts depending on very ordinary life events provide a defeater for the judgement that pre-theoretically counterintuitive desire sets are problematic.

Pereboom's argument is influenced by Harry Frankfurt's famous manipulation cases (Frankfurt, 1969). In these cases, an evil neuroscientist – call her Sarah – messes with someone's – let's say Miriam's – brain to make it the case that Miriam will kill someone else – let's say Alex. The idea is that if Miriam decides to kill Alex by herself, the action works out as usual, but if Miriam does not, Sarah's manipulations will kick in and alter Miriam's motivation so that she will desire to kill Alex and act on that desire. Hence, Miriam seems unable to do anything but killing Alex. However, independently of the manipulation, Miriam *still* decides to kill Alex, and the manipulative force is never actualized. So Miriam seems to control her action and be morally responsible for it.

Pereboom argues the other way around. He begins with a case where Miriam satisfies all mainstream conditions for moral responsibility, yet still does not seem morally responsible because she has been manipulated. (In fact, Pereboom's version of Miriam has even been created from scratch by Sarah the evil neuroscientist.) He then presents some other cases where external factors bring about an agent's actions, but the actions still seem to fit the standard conditions for moral responsibility, and concludes that the best explanation of why these cases seem to trigger intuitions suggesting that the agents involved are *not* responsible is that their actions are completely determined by background causes. It follows that we should say the same in his final case, where physical determinism holds, but standard compatibilist criteria for responsibility are satisfied. So determinism and moral responsibility are incompatible.

Whatever one thinks about this conclusion, I shall use a similar strategy to show how (non-moral) reasons co-vary with agents' desires. If we are allowed to explain slight variations in reason sets by appealing to unremarkable changes in desires, we should be allowed to explain much more normatively significant variations in reasons by appealing to changes among the desires that can explain reasons in the same way.[19] The fact that we can give such explanations shows that desire sets that, pre-theoretically, might seem to have the wrong extension to ground reasons are unproblematic. This turns the tables on the challenge of getting the extension of reasons right.

Here, we can start out with the following cases:

> (*CASE 1*) Assume that Andi has a very ordinary set of (final) desires. Andi desires to see his children, to watch his local sports team play, to go about doing his job, to participate in some sports, to read a good book, to have a drink every now and then, etc. Andi, intuitively, has reasons to do these things.
>
> (*CASE 2*) After having had too much to drink one night, Andi suddenly loses his desire to drink and becomes a teetotaller. Andi still retains all his other desires. Andi has reasons to do all these other things, but he has no reason to drink anymore.

[19] Street (2009) thinks cases of unusual desire patterns that may seem to generate weird reasons might require significant imaginative resources to make intuitively plausible. For example, one might tell some complicated evolutionary story about how someone might have evolved to attain such desires. Such stories might be right, but I think Street requires too much; the desire patterns one needs to imagine are not at all very complex, as my main argument here shows.

In *CASE 1*, Andi's desire set gets his intuitive reason set right. There does not seem to be anything problematic about his desires or reasons. In the change to *CASE 2*, however, a very ordinary life event alters Andi's desire set. This seems to be quite enough for his reasons to change along with it. There is nothing strange about thinking that one may or may not have a reason to drink alcohol depending on whether one desires to do so or not. But the kind of change that happens between Cases 1 and 2 can be generalized:

> (*CASE 3*) Due to other pedestrian life events, Andi loses his desires, one by one, to do anything else than subsisting and supporting his local sports team. For example, he wins all the tournaments he sets out to win in his sport of choice, satisfying his desire to participate in them, he retires from his career, so he no longer has a job to care about, etc. While Andi intuitively has reasons to subsist and watch his local sports team, he also seems to have *too few reasons*.

To flesh out the story, we can imagine that Andi, by now, has won his trophies, is rather old, etc. But he still desires to watch his local sports team and to subsist, and these desires still ground reasons to do those things. But that is hardly, intuitively, enough to explain the reasons agents pre-theoretically have. Andi suddenly seems to have *too few reasons*.

However, Andi has ended up in that position by what seems like a perfectly ordinary story in terms of changes in his life circumstances, just like how he lost his reason to drink in *CASE 2*. So if we can explain how he could lose his reason to drink in *CASE 2* in virtue of a shift in his desires, we should be able to explain how his reason set has diminished with his diminishing amount of desires here too. So we can defeat the intuition behind the *too few reasons* version of the *EXTENSION* objection by appealing to perfectly pedestrian desire shifts based on changes in Andi's life circumstances.

What about the *too many reasons* version of the *EXTENSION* objection? We can give an exactly analogous explanation here too:

> (*CASE 4*) Assume again that Andi has the desire set he has in Cases 1 or 2. But instead of removing some of his desires, we add a new desire to his psychology. Assume that Andi develops a taste for Indian cuisine, including its spicier dishes. Eating them is very agonizing for Andi, but he stalwarts himself and does it. After some time, however, he grows bored of the Indian spices – but not of the agony. In fact, he has developed a final desire for the agony of eating spicy food, but not one associated with Indian food in particular.

CASE 4 takes the response to the *too few reasons* problem in Cases 1-3 and extrapolates it to the problem of *too many reasons*. There is no interesting difference between adding or subtracting desires to or from someone's motivational set; both types of shifts can happen due to pedestrian changes in life circumstances. We can then see how someone's reasons, in their life circumstances, vary with their desires. So seemingly weird reasons are not so weird after all – it is not strange that someone can form desires that ground them.

This story does, in fact, generalize even beyond Cases 1-4. It is not just the case that we can undercut the bite of intuitions about the extension of our desire or reason sets if we actually go through changes in our life circumstances. Rather, from the fact that desire and reason sets with pre-theoretically counterintuitive extensions are unproblematic, we can see how these desires and reasons are not that strange in general, whether or not they have shifted like Andi's. Hence, changes in life circumstances serve as an undercutting defeater of the intuitions that suggest that there is something problematic about pre-theoretically weird desire or reason sets.

The strongest worry about this argument is, I think, that Andi's desires and reasons risk being adaptive. Many cases of desire changes seem like they might be caused by life circumstances that are in various ways normatively problematic, such as abuses of power. Does my argument legitimize such cases? Not obviously. The idealizing properties that an ideal Andi must have to explain Andi's reasons are those I assumed in section 1 above. And there is nothing in the cases that is incompatible with Andi having those properties all along, so we may assume that he has been suitably ideal all along, too.

Furthermore, many objections to Pereboom's original argument have no impact on my argument about reasons. My starting point is not an intuitively implausible case that is generalized to get counterintuitive results (*contra* McKenna, 2008), since *CASE 1* seems to get Andi's reasons extensionally right. Similarly, it is not the case that *CASE 1* Andi is not an agent, and hence not in the game for normative phenomena like reasons or responsibility (*contra* DeMetriou, 2010). And finally, it is unclear why there would be any problems with any of my cases (*contra* Fischer, 2004). So these objections are unproblematic for me. I conclude that we well may have (pre-theoretically) weird reasons.

(3) Objections (1): General Worries about Desires Internalism

The last section was no doubt long and dreary to read. But it does show that *DESIRES INTERNALISM*, featuring the properties of idealized agency, has the resources needed to explain a whole host of features of reasons that a theory of reasons should explain. Given its positive explanatory power with respect to the features of reasons mentioned, it seems very plausible. But other worries about *DESIRES INTERNALISM* – that cannot plausibly be treated as features to explain – remain. In this section, I try to clarify my view by responding to three relatively quick worries. In the next section, I engage with the conditional fallacy.[20]

First, there is a question here about which explanation is the *best* one. When I discussed *PRACTICAL SIGNIFICANCE*, for example, I argued that beliefs about reasons often play a crucial role in the formation of motivating reasons – they can trigger our capacity for instrumental rationality, or cause desires themselves. This role seems to show that the desires that explain reasons need not play a causal role in action at least independently of beliefs, so it is unclear why they would have to do any work to explain reasons or actions for reasons (cf. Enoch, 2011*b*, ch. 9; cf. Skorupski, 2010, pp. 248-253). For insofar as I act, on this view, my belief is what ultimately makes me act. It causes a desire, which then participates in Humean action-explanations.

Similarly, one may wonder whether the explanation I have given of *UNIVERSAL/INDIVIDUAL* is very plausible. I wrote that universal reasons are explained by desires that everyone is rationally required to have, whereas individual reasons should be understood in terms of more contingent desires. But is it not more elegant to say that universal reasons just are what everyone has reason to do, without invoking rationality?

However, questions about who can best explain particular data points do not generate an argument against *DESIRES INTERNALISM*. I have tried to make a cumulative case for its plausibility by appealing to its ability to explain lots of features of reasons. It follows that the view is explanatorily powerful in general, and *that* is why it is acceptable. This argument has little to do with how others may or may not be able to explain some particular phenomena; it is enough to establish that *DESIRES*

[20] There are several types of worries about desire-based forms of internalism about action that can be straightforwardly dismissed, however. First, in their Stanford Encyclopaedia of Philosophy entry on the topic, Finlay and Schroeder list three standard problems having to do with desires (Finlay and Schroeder, 2017). Potentially, desires do not explain reasons at all, or they only do so because they involve the judgement that something is a reason, or they are reason-based themselves. But these possible views are all ruled out if *DESIRES INTERNALISM* is explanatorily powerful in the way it is if the cumulative argument in section 2 is successful. For if *DESIRES INTERNALISM* is powerful in that way, it is very plausible that desires (which are not based on reasons) indeed can explain reasons (cf. appendix A for discussion).

Second, many replies to theories like *DESIRES INTERNALISM* are based on criticisms of the Humean approach to action. In particular, criticisms of Williams-style views – where (1) reasons must motivate, (2) only desires motivate, so (C) reasons must be desire-based – work like this by denying (2) (cf. e.g. Heuer, 2004). However, I have defended *HTM* in chapters 4, 5, and appendix A.

INTERNALISM can explain relevant features of reasons. Whether it, ultimately, is the best explanation of reasons for action depends on questions about metaethical assumptions that mostly are beyond the scope of this dissertation.[21]

Some other objections to internalism have appeared in exchanges between Derek Parfit and Mark Schroeder. I shall discuss two pertinent ones here.[22] The first one is Schroeder's so-called 'Wrong Place' objection (2007, pp. 37-40). The idea here is that Humean explanations of why we have reasons place the reasons in the wrong place – in our desires to help, for example, and not in the fact that someone needs help. We have, Schroeder suggests, some pre-theoretical idea about what the right explanatory structure in a theory of reasons should be. And the right place is in facts about the patient for the action, not in our desires.

The fact that desires figure in the explanations of reasons says nothing about what we have reason to do, however. Even though it may seem like it is (non-desire based) facts (or true propositions) that give us reasons, we can have reasons relations holding between such facts (or propositions) and us even though the reasons relations have their *sources* elsewhere. Rain may come from rain clouds, but it is not for that reason grounded in or composed of 'rain cloud material'. It is water, and water may, simultaneously, be given a chemical explanation as identical to H_20 molecules. It is that kind of relation that I am concerned with here. The fact that someone needs help is more like a normative analogy to the relation that holds between rain and rain clouds.

Second, there is an argument to the effect that there is something incoherent about any form of desire-based internalism (Parfit, 2011*a*, pp. 91-100; cf. Schroeder, 2007, ch. 3). If reasons have their sources in desires, we need a reason to act on our desires, and that latter reason cannot be explained in terms of our desires itself. So at least one reason is not explained in terms of desires.

But internalists need not postulate extra reasons to act on desires. As I argued in chapters 4 and 5, desires are partially *constitutive* of paradigmatic actions and agents. It makes no sense to say that we have reason to act on desires rather than anything else; rather, their explanatory role in our psychologies and actions differ from that of normative reasons to act for them.

Do we, instead, need reasons to act on any particular desires that feature in the psychologies of ideal agents, and hence to act on our idealized responses? No, those

[21] I *do* think reasons externalists tend to make awkward metaethical assumptions, however. But, queerness issues aside, I shall not endeavour to argue that here.
[22] The others are treated elsewhere in the discussion above, in particular under *EXTENSION*.

responses are *legitimized* by the fact that the ideal agent is rational, as I argued when discussing *PRESCRIPTIVITY* above. They have a kind of prescriptive force for us just in virtue of having that status. So we do not need reasons to act on our desires.

(4) Objections (2): The Conditional Fallacy

Beyond the objections to *DESIRES INTERNALISM* already discussed, probably the most significant standard worry for all forms of idealized response reasons internalism is the conditional fallacy (Shope, 1978). Generally speaking, the point of the fallacy is to point out a potential incongruency between the left-hand side of the biconditional in explanations that feature conditionals, and the relation between the antecedent and consequent in the conditional on the right-hand side (Johnson, 1999; Shope, 1978).[23] How is that a problem for me?

The core point of *DESIRES INTERNALISM* is that *A*'s reasons are explained by *A+*'s desires. But perhaps, after suitable idealization, *A+* would have desires that cannot explain *A*'s present reasons for action. If I, for example, would go through an idealization process, I may well acquire some desires that explain my reasons, while my present desires do not. But I may also acquire a new (sub)set of desires that have little bearing on what I have reason to do *now*. Ideal Olof might attain a set of desires that is irrelevant for my present situation.

Similarly, by going through the idealization process, I may lose some desires I currently have. This, too, is sometimes helpful, as I may have problematic desires. But sometimes, I may lose too many desires, leading my idealized counterpart to lack the desires needed to explain my present reasons. If either the scenario where I gain desires that cannot explain my present reasons or the scenario where I lose desires so my ideal desires cannot explain my present reasons occurs, it seems like the relation between the antecedent and the consequent on the right hand-side of *DESIRES INTERNALISM* is

[23] Here is the original statement of the fallacy: '[the conditional fallacy is a] mistake one makes in analyzing, defining, or paraphrasing a statement *p* or in giving necessary and sufficient conditions for the truth of that statement, by presenting its truth as dependent, in at least some specified situations, upon the truth (falsity) of a subjunctive conditional, *C*, of the form: 'if state of affairs *a* were to occur then state of affairs *b* would occur', when (…) [one] has overlooked the fact that in some of the specified situations the occurrence of certain relations involving factors that are mentioned in *p* or in the *analysans* (*definiens*, paraphrase, or list of necessary and sufficient conditions) is connected either with the occurrence of *a* or with the absence of *a* in such a way as to be responsible for a disparity between the truth value of *p* and the truth value of the *analysans* (*definiens*, etc.) in those situations' (Shope, 1978, p. 403, slightly editorially adjusted)

incompatible with the reasons I actually have. The idealization alters my desire set, making it different from one that may explain my reasons.

The conditional fallacy, then, seems to be a version of the *EXTENSION* objection. *DESIRES INTERNALISM* risks introducing reasons that we do not have or rule out reasons that we do have, hence generating several cases where it seems to fail to get the extension of our reason sets right. And what is worse, the problem remains despite my Four Cases response to *EXTENSION*, because the intuitive worry that we have not got agents' reasons right in at least some cases reappears. My defeater response to *EXTENSION* shows how we well may have pre-theoretically weird desire or reason sets, but that does not really speak to the issue of whether there might be some sort of problematic *incongruencies* between the *A*'s and *A+*'s desires. It is one thing to say that there can be very divergent desire or reasons sets, and quite another to say that *A* and *A+* seem to stand in awkward relations to each other.

But does *DESIRES INTERNALISM* really suffer from the conditional fallacy, or is there a plausible response to it? I will start answering this question by discussing the first type of case above, i.e. where *A+* has acquired a desire that *A* presently lacks, but which is irrelevant for explaining *A*'s present reasons. Here is a well-known version of it:

> (*SQUASH*) You lose a good game of squash. This makes you very angry, and your anger gives you a desire to punch your opponent. You know that you should walk up to her and thank her for a good game, but you also know that if you were to walk up to her to thank her for a good game, you would act on your desire to punch her. Hence, it might seem like you have more reason to walk away from the game than to thank her. What do you *really* have reason to do?[24]

Assume that the desire to thank your opponent is such that you would acquire it if you were to go through an idealization process, but you do not currently have it. If so, in *SQUASH*, the reason you have which is grounded in this desire plausibly trumps your reason based on an anger-based desire to punch the opponent. But, as of now, you have a strong desire to punch your opponent if you were to walk up to her, and *would* act on that desire if you were to do so. Hence, intuitively, you may seem to have more reason to walk away from your opponent than to act on the reason grounded in your idealized counterpart's desire. How should we interpret the reasons you have in this case?

[24] This case comes from Smith (1995), but see also (Henning, 2018, ch. 5; 6; Millgram, 1996).

According to Korsgaard (1986), what you have reason to do is what your ideal counterpart would do – the counterpart serves as an *exemplar* that you should try to emulate. So what you have reason to do in *SQUASH* is to walk up and thank your opponent. True, if you were to walk up to your opponent, you would still punch her rather than thank her, but that is beside the point. What you have *reason* to do is what the ideal agent desires (and would do).

On Smith's (1995) view, however, your idealized counterpart in world *w** serves as an *advisor* to non-ideal you here in world *w*. Here, what you have reason to do (in *w*) is what your counterpart would advise you to do, presumably in the light of both your and her desires. On this view, your counterpart would plausibly advise you to walk away, and this piece of advice is what serves to ground your reason to walk away. This seems like the intuitively right result. Hence, the advisor view seems better suited than the exemplar view to explain your reason(s) in cases where the ideal agent has acquired a desire that you, as a non-ideal agent, lack.

What about the version of the conditional fallacy where *A*+ loses some of *A*'s desires in the idealization process? We can re-interpret *SQUASH* to show how this works. Assume that your two strongest desires when the game is over are a desire to punch your opponent and a desire to thank her. But assume also that your idealized counterpart *loses* her anger-based desire to punch the opponent in virtue of her idealization. It follows that her strongest desire is to shake her opponent's hand, and hence you plausibly have a reason to do so. For another example, if I were to desire to spend all my money on gambling, then I might have reason to get rid of that desire. But an ideal version of me would not, intuitively, have that desire in the first place. It would be lost in idealization (cf. Henning, 2018, ch. 5-6; Johnson, 1999; Millgram, 1996). What do you (or I) have reason to do in these cases?

On an advisor view, explaining these cases is pretty straightforward. Ideal advisors would, plausibly, advise us to avoid walking up to our opponents or to gamble. And they would do so even though – or, possibly, just because – they lack the desires to punch our opponents or to gamble.

The exemplar view does worse here. The non-ideal agent's reasons are supposed to depend on the desires of the ideal agent – but since the ideal agent lacks the problematic desires, she does not need desires to avoid getting into trouble. It is perfectly possible to be an ideal agent without desires to get rid of a desire to punch opponents or to gamble; idealized agents, as I understand them, may have all kinds of desire sets, so they do not *need* those desires – and especially not so in cases where they lack the problematic desires they might want to get rid of. So both when the ideal agent acquires a new desire or loses an old one in idealization, the advisor view does better than the exemplar view. We therefore get a *prima facie* case for the advisor view. I shall spend the rest of this section refining it in the light of the many objections that have been raised against it.

One objection to the advisor view that actually helps us develop it comes from Bedke (2010). He argues that it is unclear what desire a fully idealized agent would have to give advice to a non-ideal one. But we can just stipulate that the advice *A*+ gives is the advice *A*+ would give to *A* had *A*+ been concerned with giving advice to *A* (that matters to *A*). There are many different ways to explicate why *A*+ would have this concern. Perhaps *A*+ just has a desire to do so, or *A*+ identifies with *A*, viewing *A* as an extension of herself and hence *A*'s actions as matters she is prudentially concerned with, or something else. The details of *A*+'s interest in *A* do not really matter; the idealized agent is a theoretical tool, not an actual agent whose desires may vary independently of her theoretical role.

Is there a risk that the advice 'that matters to *A*' here in fact requires that *A*+ responds to external reasons, or at least is set up in such a way that we smuggle externalist assumptions about what we have reason to do in *SQUASH* (and similar cases) into *DESIRES INTERNALISM*? Not really. We can interpret the advice that matters to *A* as advice that is pertinent for establishing what the thing to do is, whatever it is, for *A* in her circumstances (cf. chapter 5, section 2; appendix B, section 1). What 'the thing to do' is is not by itself to be understood in terms of reasons; it is a catch-all concept for what one is to do in any given situation, whatever normative currency we are trading in, so this move should be theoretically safe to make.

But does explicating reasons in terms of advice about what matters to *A*, construed as what is pertinent for establishing what the thing to do is, still risk making us appeal to some sort of reasons externalism deeper down? The worry is that the advice *A*+ will give *A* just requires us to appeal to what *A* seems to have reason to do, independently of her desires.

In response, I shall construct a version of the advisor model that does not. I shall not endeavour to conclusively establish how the relation between desires and advice looks here, however. It will enough to present a toy model that does not rely on external reasons and gives the right result in the *SQUASH* cases. Suitably developed, I shall assume that it, or something like it, holds generally.

Assume, then, that *A*'s final reasons are explained by *A+*'s desires. Then there are two options. Either *A* shares *A+*'s desires and is in similar circumstances *C* as *A+*, or she does not. If she does, *A+* would plausibly just advise *A* to act on the desires they share. In these cases, *A*'s desires coincide with her ideal desires. To recycle an example from chapter 3, section 6: if *A+* prefers spaghetti Bolognese to lasagna, and *A* does too, and all background circumstances are equal, *A+* will just advise *A* to have spaghetti Bolognese rather than lasagna.

However, there can also be cases where the desires or circumstances of *A* and *A+* come apart. Here, we can construe *A+* as advising *A* to act instrumentally so that *A* will satisfy *A+*'s weightier desires (on some suitable conception of weight) as well as possible given *A*'s desires and circumstances. There are no external reasons invoked here, for *A+*'s desires are just the desires *A* would have after due idealization, and the model gives the right results in the *SQUASH* cases.

How? If *A+* has acquired a desire to thank *A*'s opponent in virtue of the idealization process – and this desire is weightier than the desire to punch the opponent, which it presumably is if the idealization process is at all functional – then *A+* would advise *A* to walk away rather than punch the opponent, since that is a better way for *A* to satisfy *A+*'s desire to thank the opponent. Perhaps *A* can thank the opponent later. Alternatively, maybe some other desire of *A+*'s is more important than the desire to thank the opponent, so *A* should walk away to act on that desire.

On the other hand, if *A+* has lost the desire to punch the opponent by being idealized, so the desire to thank *A*'s opponent is her strongest desire, it might seem like *A* now has a strong reason to just walk up to the opponent to thank her. But because of the circumstances *A* is in, *A+* will not just advise *A* to act on that desire, for that would just have *A* punching the opponent. Hence, *A+* would presumably advise *A* to walk away in the same way, and with the same rationale, as in the acquired desire case.

By now, I have been able to develop the advisor model by responding to Bedke (2010). But the worry that the view still might fall back into some form of externalism remains. In his (2000), Eric Wiland argues that the kind of advice we actually receive on the advisor view pulls the view closer to externalism about practical reasons. This is because the advisee need not share the advisor's desires about what to do. Hence, the reason given by the advisor's advice is in one sense external to the advisee's motivations.

But this worry relies on the stronger, Williams-style, connection between reasons and motivation that I explicitly refused to assume when I discussed *PRACTICAL*

SIGNIFICANCE. It does not matter for my (or Smith's) conception of motivation by reasons that non-ideal agents do not have the desires that ideal agents have. They can form new desires by recognizing their reasons, as per the discussion of *PRACTICAL SIGNIFICANCE*.

Another problem for the advisor view that Wiland has emphasized stems from our actual advice-giving practices (Wiland, 2003). He argues that, in our actual advice-giving practices, advice can sometimes be based on the assumption that the advisee is supposed to ignore the advice – and will do better *because* she ignores it. This means that the advisor view seems incompatible with our advice-giving practices.

However, we can just stipulate that the advice given in the theory of reasons is honest, much like we could stipulate that the advisor is willing to give advice in response to Bedke's worry. If the advice is honest, it can plausibly explain our reasons, even though our actual advice-giving practices might be more intricate than the advisor model. But that does not matter; again, the advisor is a theoretical tool, not an actual agent.

Maybe most famously, however, Robert Johnson has tried to argue in several papers that the advisor view is implausible (Johnson, 1997; 1999; 2003). Johnson claims that:

> [T]o the extent that [Smith's] model connects reasons to advice, it is not a model of the internalism requirement [i.e. the claim that reasons must be able to motivate action] at all. Yet, to the extent that it connects reasons to motivation, his model collapses into the [exemplar] model. (Johnson, 1997, p. 619)

Johnson's point is that the advisor view loses track of the relation between reasons and motivation because it no longer can explain the connection between them. According to *DESIRES INTERNALISM*, if we were to be fully ideal, we would have the same desires as our idealized counterparts because they are just who we would be if we were ideal. However, those desires need not play an explanatory role when we act for our reasons on the advisor view. If we are to be motivated by what our ideal advisors say, we would be responding to their advice rather than acting on our idealized desires. They need not desire to act on that advice themselves, and we need not have the desire that features in the advice either. So their advice has little to do with what we would be motivated to do if we were fully ideal.

On the other hand, Johnson thinks that to the extent that what we believe their advice would be – i.e. what our reasons would be – in fact might motivate us, that is because *that* is what we would ideally desire, and we already have that desire. But that is exactly how motivation works on the exemplar model. So the advisor model collapses into the exemplar model.

I think this argument goes wrong on both fronts. Starting with the point that the advisor view risks falling back into the exemplar view, it seems unfair to Smith to characterize his view as 'falling back' into the exemplar model. The ideal agent can often give advice based on what she desires for herself, like in my pasta case above. Here, the advisor view works like the exemplar view. But sometimes, the advice is more complicated, taking the non-ideal agent's imperfections into account, and I tried to show how that, too, might work when I discussed the *SQUASH* case in response to Bedke's worry. So it is not the case that the advisor view falls back into the exemplar view – rather, it extends it.

To go back to the first point about reasons and motivation instead, reasons can still motivate on Smith's advisor view. As I argued in response to Wiland's point about how the ideal advisor view might collapse into externalism, Smith and I hold that beliefs about the reasons we have can *cause* desires, instrumental or final, for agents. There is no need to think of 'the internalist requirement' in the way Williams does.

In fact, even Johnson (1999) himself concedes that Smith thinks this way, though he discusses Smith's actual position as a final possible response to his original mischaracterized version of it. There, he also replies to the position by appealing to a conditional fallacy worry which is equivalent to the second version of *SQUASH* case, where the ideal advisor loses some desire when idealized. But, as I have argued, the advisor model can handle that case.

Hence, it seems like the advisor model survives Bedke's, Wiland's, and Johnson's worries. But there are more worries still. First, I have treated the conditional fallacy as a worry about the extension of our reason sets. But it is not obvious that the advisor view gets the extension of our reasons right. Return to *SQUASH*. If the angry squash player had been perfectly rational, her strongest reason would not have been to walk away, but to thank her opponent. This means that the advisor view might seem to do away with our strongest reason because it makes our strongest reason conditional on our imperfections. Is that extensionally awkward?

Not necessarily. Perhaps our imperfections do have impact on the reasons we have in any set of circumstances *C*. For remember that *DESIRES INTERNALISM* indeed is supposed to be a theory about which reasons *A* has in circumstances *C*. The circumstances can plausibly include *A*'s imperfections. It is unclear why we would not allow them to matter.[25]

The second remaining worry is that we may have lost track of some of the other features of reasons from the positive argument for *DESIRES INTERNALISM* in section 2. One may wonder whether the advisor model will make it possible for us to be so alienated from our reasons that we no longer can explain *RELATIONAL CHARACTER*, for example. Intuitively, reasons that are not closely related to an agent's present desires are no longer her own. If so, the difference between the idealizing conditions and our own world might make it hard to show why our reasons are ours.

However, I argued that what explains *RELATIONAL CHARACTER* is that the desires that explain reasons are the desires of ideal versions of agents and that we easily can be motivated by them, not just that they are close to our present desires. That sometimes happens, making it easy to be motivated by our reasons, but our reasons need not always be closely connected to our present desires. So there is no tension there.

Another worry is that we may have lost track of *KNOWLEDGE*. I argued that there were several ways in which we may come to know our reasons. We may know them by knowing our present desires, knowing general facts about what desires people tend to have in their circumstances, through Railton's feedback mechanism, and sometimes by transcendental reflection. On the other hand, insofar as the ideal advisor's advice may deviate from what we desire, it is less clear how we can come to know what our reasons are. Presumably, the transcendental argument will look the same regardless of how advice might shape our reasons, but what are we to say about our knowledge of our individual reasons?

There is nothing that makes the advisor model worse off than the exemplar model here, however. The advisor will presumably usually give advice that is close to the agent's own desires, and there is nothing that would make the exemplar model a more appropriate theory about our reasons in the light of environmental feedback than the advisor model. (Indeed, Railton's model of the good is formulated as an ideal advisor theory in the first place.) There are no extra difficulties about advice here, and as we plausibly can attain knowledge about our reasons in several ways, we can plausibly do so even if they are advice-based.

[25] If we think like this, perhaps we can even treat *A*'s reason to thank her opponent as a conditional reason that does not apply in her present circumstances *C*. But if *A*'s motivation would change, the reason might become applicable.

To summarize: I have defended the advisor model in response to the conditional fallacy. It seems well-suited to get the extension of our reasons right, and it is able to respond to many challenges. Given the explications of the view, what *A* has reason to do according to *DESIRES INTERNALISM* is what *A*'s idealized counterpart *A*+ in *w** would advise *A* to do in *w*, (*i*) where the advice is what *A*+ would have been concerned with giving had she been concerned with giving advice to *A* (that matters to *A*), (*ii*) is what *A*+ desires for herself if her and *A*'s desires coincide and they share similar circumstances, or is what would satisfy *A*+'s weightier desires instrumentally as well as possible if their desires or circumstances come apart, and (*iii*) *A*+ is honest.

(5) Conclusion

In this chapter, I have tried to show that *DESIRES INTERNALISM* is a plausible theory about our reasons for action. In section 1, I set up the view, and in section 2, I showed how it can explain many key features of reasons. Then, in section 3, I replied to some general objections to internalism, and in section 4, I defended an advisor interpretation of the view. In the next chapter, I shall extend it to cover moral reasons.

7. Morality as Cooperation

Most of the constructive arguments in this dissertation have been presented by now. In chapter 4, I presented an argument for a Humean theory of agency based on a version of *HTM*. In chapter 5, I argued that *HTM* should be supplemented with a principle of instrumental rationality (and that the principle is normative). I have also started to develop a two-tiered form of constitutivism. The first tier involves the instrumental principle, and the first step of the second tier involves defending an internalist theory of practical reasons (in chapter 6).

But I have yet to explain moral normativity, which is what I fundamentally set out to do in this dissertation. Furthermore, making sense of (universal) moral reasons is also crucial for concluding the argument in chapter 6. Such reasons featured in the explanations of *KNOWLEDGE*, *UNIVERSAL/INDIVIDUAL*, and in the *too few reasons* version of the *EXTENSION* objection.

Hence, to wrap up the main argument of the dissertation as well as the argument for *DESIRES INTERNALISM*, even though this move not is an orthodox one for Humeans, I shall now attempt to defend some (universal) moral norms.[1] To that end, my strategy takes hints from Michael Smith's (cf. chapter 2, section 5). He argues that some desires are constitutive of idealized versions of us *qua* agents, and because the desires of idealized agents explain reasons, it follows that we all have reasons based on the desires that they have.

But while I have criticized some of Smith's moves, I shall now try to improve on them. I shall argue that idealized agents must have some idealizing properties from which it follows that they must have desires to cooperate with other cooperative agents to satisfy their other, respective, desires.[2] By the reasons internalist assumption that their desires explain our reasons, such pro-cooperative desires can then explain at least some moral reasons. Simultaneously, the argument will show that idealized agents are required not to have certain anti-social desires. It follows that we cannot have reasons for action stemming from such desires either.

Moral norms based on these properties of idealized agents can be captured in my constitutivist framework. Using the abbreviations from chapter 1, on my account (*T*),

[1] Some recent Humean attempts are, however, (Dorsey, 2018; Driver, 2016; Manne, 2016; Schroeder, 2007; cf. chapter 2, section 1, for discussion).

[2] What is 'cooperation'? Good question. I shall only assume that cooperation involves several agents trying to achieve some end together (cf. Regan, 1980, p. 129). The reader is free to fill in with more.

there are two fundamental moral norms (P_{PM}). First, there is a universally prescriptive reason to cooperate, grounded in desires that are partially constitutive of idealized agents, i.e. C of A^*. Second, another condition of the ideal agent is that she cannot have anti-social desires. These fundamental norms shape the way other moral reasons look in a full theory of moral reasons.

To get there, however, in section 1, I explicate the conceptual framework that I will use to discuss moral reasons. In section 2, I present the argument from idealization, which suggests that a final desire to cooperate with other cooperative agents is partially constitutive of ideal agency. In section 3, I show how that argument can be extended to explain moral norms. In section 4, I discuss objections – including the shmagency objection. I conclude in section 5.

(1) Conceptual Background

I shall start off by clarifying my strategy. I mentioned in chapter 1, section 2, that I primarily aim to develop a constitutivist theory about moral norms. Moral norms, I take it, are universally prescriptive and conventionally recognizable as moral.

What does universal prescriptivity mean? I argued in chapter 1, section 2, that by this type of prescriptivity, some normative phenomenon must have normative force for all agents (at least if they are, at all, minimally sophisticated). This is true whether we are talking about reasons or other norms. Moreover, to show that the moral norms I aim to explain indeed are moral, they must be conventionally recognizable as moral. I shall attempt to establish at least two central norms that are like that.

However, while the view I shall propose involves two central moral norms, it is simultaneously a hybrid view about moral reasons. It features three kinds of them. First, there is a universally prescriptive and morally recognizable reason to cooperate with others. Second, there are *non*-universal, i.e. individual, reasons. Some of the latter may be recognizably moral. In this case, I shall call them *moral ideals* – though note that such reasons are only moral reasons in secondary sense of the word, as they are not universally prescriptive. If they are not recognizable as moral, however, I shall call them *practical interests*.[3]

[3] There is, presumably, also a grey area somewhere between them, making it unclear how we should characterize some reasons. This is not a problem; there are almost always hard cases.

Third, I shall introduce a special kind of universally prescriptive reason which is more contingent than the reason to cooperate. This kind of reason involves the reasons one must act on to be able to cooperate via the reason to cooperate. One will be structurally required to act on them insofar as one is involved in a cooperative situation, viz. where agents cooperate together. However, since agents' desires vary, there will be different things that agents will cooperate to achieve, depending on their varying desires.

Why these three kinds of reasons? I will argue that one of the two fundamental moral norms that my view generates is one according to which we – roughly – have reason to act cooperatively (with other co-operators) to satisfy our other respective desires. According to the other one, anti-social desires or reasons are ruled out. But then, in the light of *DESIRES INTERNALISM*, the other desires that can go into cooperation will be the sources of moral ideals or practical interests. And we may satisfy these desires by acting cooperatively via some sort of shared arrangements, hence generating a third kind of reasons. Accordingly, the overarching norms suggesting that we cooperate non-antisocially will co-generate what we actually have reason to do together with our other desire-based reasons, be they moral ideals or practical interests. But if we are in the right circumstances, it is necessary that we do so in *some* way.

It is notable that this view possibly may turn out to have mildly revisionary consequences for how morality looks (cf. chapter 1, footnote 8). But I do not think the moral notions I am about to explain exhaust the moral sphere. I will not say anything about obligations, rights, duties, supererogation, suberogation, values, or even about whether even more properties should be added to the ideal agent here. It might – but need not – be the case that many of these notions can be captured when my view is developed to be more comprehensive. But for now, I shall only focus on the central ideas in my account. How they relate to other first-order notions is an issue for another time.[4]

(2) The Argument from Idealization

In this section, I shall present and defend an argument for thinking that the idealized agent who features in *DESIRES INTERNALISM* must have a desire to cooperate with other cooperative agents. In later sections, I intend to explore the normative implications of this desire. Here is the argument:

[4] One might, however, helpfully compare the way I explain morality here with Scanlon's (1998) insistence that he is only after explaining what we owe to each other, not necessarily all of morality.

(1) If *A*+'s psychology is able to explain the reasons of an agent *A* in our world, then *A*+ is suitably idealized.
(2) If *A*+ is suitably idealized, then *A*+ has a set of idealized desires (based on *A*'s desires) for what to do in a range of situations or circumstances, many of which feature the circumstances of justice.
(3) If *A*+'s psychology is able to explain the reasons of an agent *A* in our world, then *A*+ has a set of idealized desires (based on *A*'s desires) for what to do in a range of situations or circumstances, many of which feature the circumstances of justice.
(4) If *A*+ has a set of idealized desires (based on *A*'s desires) for what to do in a range of situations or circumstances, many of which feature the circumstances of justice, *A*+ must have a desire to cooperate with other cooperative agents as a matter of being suitably idealized.
(5) If *A*+ must have a desire to cooperate with other cooperative agents as a matter of being suitably idealized, *A*+ has a desire to cooperate with other cooperative agents as a partially constitutive feature of her idealized psychology and herself.

(C) If *A*+'s psychology is able to explain the reasons of an agent *A* in our world, *A*+ has a desire to cooperate with other cooperative agents as a partially constitutive feature of her idealized psychology and herself.

To start off, premise (1) might seem fairly obvious by now. I defended the need for idealization to explain reasons in chapter 6. Idealization does a lot of work there; I argued that it makes sure that the agent has the right desires to explain reasons because it ensures us that the ideal agent is fully functional or fully manifests her capacities insofar as she acts, plus that she is in the relevant background conditions for doing so. So premise (1) seems safe.

Nevertheless, suitable idealization has some important implications that I shall introduce here. They, in turn, have surprising normative upshots. To get to those implications, I shall start off with some recapitulation. In chapter 2, section 5, I argued against Michael Smith's take on idealization. Smith argues that in cases where our agential capacities to satisfy our desires and hold true beliefs may conflict, coherence-inducing desires can hinder them from clashing and help us in their exercise. These desires, he thinks, generalize, so we must desire to help and not interfere with others' exercise of their capacities as well. By reasons internalism, it follows that we have reason to help and (otherwise) not interfere with others' use of their capacities.

I argued, however, that Smith has not presented a reason to think that an ideal agent risks running into situations where her capacities to realize her desires and believe truly conflict. Such situations can be avoided in other ways than by attributing desires to help and not interfere to ideal agents. Perhaps one aspect of rationality is a principle of coherence whereby the ideal agent is sensitive to capacity conflicts and hence avoids them. Or perhaps she is partially constituted by a Davidsonian background principle that guarantees that she cannot have an incoherent psychology, or maybe she must be coherent to explain reasons as a matter of idealization. Since Smith has not ruled out these alternative options, ideal agents need not have coherence-inducing desires, so he cannot get the moral reasons he wants out of his ideal agents.

This discussion does, however, open up a question. If Smith's agents do not risk becoming incoherent in their circumstances, they do not need reason-explaining desires. But we may wonder what circumstances an ideal agent may be in or have desires about more generally. By 'circumstances', I mean the natural, social, physiological or psychological background conditions that an agent faces or could face, viz. the conditions of those kinds that may affect her psychology, decisions, or actions. Examples of what I have in mind are what species she belongs to, which planet she inhabits, and what society she lives in.

To introduce some more terminology, when an agent is in a particular set of circumstances, I shall call this a *situation.* As an ideal agent can plausibly be in or have desires for what to do many situations, situations may sometimes be understood as possible worlds, however they should be interpreted if they are to be compatible with interpretations that do not have any controversial ontological implications about what they or the ideal agent must be like.[5] For the same reason, a situation may sometimes be a subset of the circumstances inside some world – an ideal agent can be in or have desires about what to do in many possibly subsets of circumstances there too. However, if you do not like possible worlds-talk, feel free to reinterpret what a situation is using your preferred interpretation of 'sets of circumstances'. Fundamentally, what matters here is that the ideal agent may inhabit or have desires about what to do in different natural, social, physiological and psychological circumstances. This has important ramifications.

Why? As internalists typically do, I have argued that agents' idealized psychologies explain our reasons. But *DESIRES INTERNALISM* does not, by itself, say which

[5] Of course, if the reader prefers more ontologically heavy-duty possible worlds, she should feel free to go with them instead.

circumstances *A*+ must be in or have desires about what to do in, only that *A*+'s desires explain the reasons *A* has in *A*'s circumstances *C*. This seems to make it possible that *A*+'s situation (or the situations she has desires for what to do in) may differ from *A*'s.

But too great divergences between *A*+'s situation or desires and *A*'s circumstances could, in turn, give *A*+ different kinds of desires than those *A* plausibly has – and therefore give *A* different reasons than those she plausibly has. That would decrease the explanatory power of *DESIRES INTERNALISM*: for example, it is harder to explain *KNOWLEDGE* if ideal agents' desires are too different from ours, for then it is harder to learn what we have reason to do through our desires.

Fortunately, this worry can be solved by resources internal to my theory. Idealization is supposed to have two main dimensions: *A*+ is supposed to be a fully functional (or a fully capacity-manifesting paradigmatic agent when she acts) in background conditions relevant for maintaining her full functionality. Hence, *A*+'s psychology (or circumstances, which might alter her psychology) should not plausibly be altered in more ways than by idealization in these two dimensions, for further changes are irrelevant. With that point in mind, something like the following principle becomes a constraint on any idealized agent whose desires explain our reasons:

> (*CLOSE*) *A*+'s idealized desires explain *A*'s reasons only if *A*+'s desires and other psychological states range over circumstances that are similar enough to those *A* may be in.

This means that at least some circumstances that *A*+ inhabits or has desires for what to do in must be similar to those *A* is in.[6] There are hard questions to ask about how we should understand that similarity (e.g. in terms of possible worlds in some technical sense?), but while such questions are interesting, taking positions on them would risk making controversial commitments that do not matter for present purposes.

Instead, here, it is enough to have an intuitive grasp on the limits *CLOSE* sets: *A*+'s desires must range over circumstances that are similar enough to those *A* is in if they are to explain *A*'s reasons. Hence, for example, if *A* is a human, we can safely rule out cases such as when *A*+ is Cthulhu from explaining *A*'s reasons. Cthulhu's desires, let alone background natural, social, physiological or psychological conditions, are plausibly

[6] Is *A*+ not a *hypothetical* agent, and hence unable to 'inhabit' situations where her desires might change? Well, if we can think of hypothetical agents, we may also think of the habitation hypothetically.

very different from those of any human. Understood like that, a condition like *CLOSE* seems extremely plausible.

There is also another, similar, property of *A*+'s that does very important work when it comes to explaining *A*'s reasons. This property is:

> (*ROBUST*) *A*+ must have psychological dispositions and capacities that remain the same over minor changes in the circumstances she may inhabit or otherwise have desires for what to do in.

ROBUST, too, can be explained by idealization. Idealization involves making a paradigmatic agent fully functional or fully manifesting her capacities insofar as she acts, as well as making sure that she is in the right background conditions. *ROBUST* is such a background condition. This is because *A*+ hardly can manifest her psychological states to act if they were to change capriciously with various more or less randomly occurring events – and this would soon also undermine their functionality. If *A*+'s desire to drink when thirsty were to turn into a desire to wear a red jumper when thirsty because her neighbours have acquired a cat, she would not be able to act on the desire to drink if her neighbour indeed did acquire a cat. Then she would soon die of thirst, completely undermining the functionality of her psychology. A 'minor' change, then, is a change in *A*+'s circumstances which is such that, if *A*+ had been sensitive to it, it would undermine her being ideal. *A*+'s psychology must be *ROBUST* in the face of such changes.

How do *CLOSE* and *ROBUST* interact with Smith's argument? He often talks about how an agent must be able to exercise her rational capacities robustly, i.e. over changes in their circumstances, including their possibly being undermined, so he seems to assume that something like *ROBUST* holds (cf. Smith, 2012*a*). There we agree. However, I argued in chapter 2, section 5 – and recapitulated above – that Smith cannot plausibly get his desires to help and not interfere out of *ROBUST*.

But perhaps *CLOSE* resuscitates his argument? Maybe the reason that the ideal agent needs desires to help and not interfere to make sure that her rational faculties function over time is that, because those capacities are like ours in situations like ours, they genuinely risk stopping to function. This would provide a way to supersede the alternative coherence-inducing properties of the ideal agent I argued that Smith has not ruled out, for it might also be argued that the agent is unlikely to be able to follow those alternatives in circumstances similar to ours. Coherence-inducing desires might preserve an agent's ideal functioning better than the alternative capacities I have suggested.

However, it is unclear why desires would do the relevant job better than some other psychological features. Desires do no special work that other mental states could not do; in particular, if the ideal agent ought to have a rational capacity or principle whereby she is sensitive to incoherence and prepared to avoid it, that capacity should also be able to apply in the relevant circumstances in exactly the same way as a desire. So *CLOSE* does not rehabilitate his argument.

On to premise (2). I assume that the circumstances of justice include the sort of things that Rawls (1971, pp. 126-130), following Hume (1978, pp. 473-534), took for granted to apply in ordinary human circumstances.[7] The most significant one is that people's desires usually cannot all be easily satisfied given the constraints that their social circumstances put on them, whether these desires are materialistic or idealistic.

Moreover, in the circumstances of justice, there is a moderate scarcity of resources, moderate generosity on part of others (or moderate ideological agreement between agents), and it is within others' power to – either individually or together – thwart any given individual's attempts to satisfy her desires by overpowering her. (To be clear, this use of power need not be moral or nice; the point is that agents are able to use their power to harm each other.) Under these circumstances, living in cooperative societies usually benefits individual agents, but participating in them does not always lead to the best results for any individual agent, given what they desire.

CLOSE ensures us that ideal agents must have desires about what to do in an extensive set of situations that feature these circumstances of justice. To recapitulate, *CLOSE* says that *A+*'s idealized desires explain *A*'s reasons only if *A+*'s desires and other psychological states range over circumstances that are similar enough to those *A* may be in. But our situation contains the circumstances of justice, and situations that do not would be very different from ours just because they would not feature the circumstances of justice. Moreover, there may be all kinds of differences between different versions of the circumstances of justice that we inhabit. There are already many such versions in the actual world, and there may be further ones still. So if *A+* had lacked

[7] Rawls distinguishes between subjective and objective circumstances, where the subjective circumstances include differences in values and life plans among people as well as ordinary human flaws, but the objective circumstances have to do with external features of the world (e.g. that there is a moderate scarcity of resources). Hume more clearly emphasizes how hard it is to satisfy individual wants and needs given moderate scarcity, moderate generosity, and the possible destruction of us and what we hold valuable by others. But they are fundamentally after the same points.

desires for what to do over an extensive set of such circumstances that we may inhabit, *A*+'s desires would not range over circumstances that are similar enough to ours to explain our reasons.

True, it is also possible that some agents do not inhabit these circumstances, or if they presently inhabit them, they may come to leave them. Hence, ideal agents should also have desires about what to do in at least some *other* circumstances. But any even remotely humanlike creature is also likely to risk being in the circumstances of justice, so their ideal counterparts should have desires about what to do in situations that feature them, too. This means that premise (2) is in place. And premise (3) follows.

Premise (4) says that if *A*+ has a set of idealized desires (based on *A*'s desires) for what to do in a range of situations or circumstances, many of which feature the circumstances of justice, *A*+ must have a desire to cooperate with other cooperative agents as a matter of being suitably idealized. This conclusion follows from the nature of idealization and *ROBUST*.

How? I shall argue that because the ideal agent has desires for what to do in an extensive range of situations featuring the circumstances of justice, to be able to exercise her instrumental rationality in a *ROBUST* way, she must have that cooperative desire (in all such situations), for that desire is what allows her to be instrumentally rational in a *ROBUST* way. And *ROBUST*, I have argued, follows from *DESIRES INTERNALISM*, so the 'must' here is not normative. It is explained by the features of idealization.

Now, it is well known that, *prima facie*, it need not always be better for individually self-interested agents to abide by the rules of justice. Gyges, Fooles, and Sensible Knaves populate the history of philosophy. These characters are sometimes better at satisfying their desires than the virtuous are. Nevertheless, having desires for what to do in the circumstances of justice, *A*+ benefits from participating in human societies, including benefiting just from living in a society in general.

In fact, being able to enjoy the good of cooperation is a matter of *A*+'s idealized instrumental rationality. By IR_{HUMEAN}, I have taken instrumental rationality to involve taking the best means one believes there are to one's ends. And the two main idealizing conditions of the ideal agent are that they, first, are supposed to have the features constitutive of paradigmatic agency fully functional or manifested when acting, and, second, that their background conditions are the right ones for their functionality.

Now, to be able to be fully functional or manifest her capacity for instrumental rationality in paradigmatic actions, an ideal agent must be able to take the best means she believes there are to her ends (cf. chapter 5, section 2). This is because paradigmatic action involves doing that. Similarly, the ideal agent must be able to take the relevant means to satisfy some different desires, since the agent can have multiple, conflicting or changing, desires. This means that insofar as the ideal agent has desires for what to do in the circumstances of justice, the goods of social interaction are necessary background conditions to ensure that her psychology is functional or possible to manifest when acting. And this is because, in situations featuring the circumstances of justice, social interaction generates more and better means both relative to the ideal agent's existing desires and relative to other possible desires she may have. The former is usually the case, and the latter is always the case, for even if agents do not have desires that are better satisfied using means available in social interaction, they can always acquire such desires. So in the circumstances of justice, social interaction is a background condition for the ideal agent's instrumental rationality.

Then we may draw a distinction. Either the ideal agent has a final desire to engage in cooperative schemes, at least given that other agents also do so, or not. If she does, all is well with her when she participates in social interaction. She will happily do so. But assume instead that she lacks such a desire. She need not necessarily be a disinterested maximizer like Gyges, the Foole, or the Knave; she can be anyone who doubts the value of any kind of cooperative arrangements but still benefits from them. What matters is that she lacks the final desire for cooperation.

If $A+$ lacks that desire to cooperate, however, but still benefits – again *ex hypothesi* – from those schemes with respect to her instrumental rationality, then she is essentially a free rider. Free riders will, in many situations, be punished by the other participants in the cooperative schemes. In the extensive range of situations featuring the circumstances of justice for which $A+$ must have desires, there are no doubt some where that happens.[8] However, in situations where free riders would be punished, an agent with a final desire to cooperate would be able to be more fully instrumentally rational, whereas an otherwise ideal agent who lacks that desire would not. The other agents would not, plausibly, punish an agent with a desire to cooperate.

But then comes the magic trick. An ideal agent who lacks the desire to cooperate is not able to be fully instrumentally rational with respect to taking the best means to

[8] Note that the ideal agent here would be a free rider with respect to the means she can take to her ends. That should be enough to run the argument, for I presume that making use of possibilities that other agents' work give her is enough to annoy some others. But it should also be possible to run the argument, *mutatis mutandis*, with the agent free-riding with respect to her desire satisfaction.

satisfy her desires in a *ROBUST* way. *ROBUST*, I wrote, says that *A*+ must have psychological dispositions and capacities that remain the same over minor changes in the circumstances she may inhabit or otherwise have desires for what to do in., where a 'minor' change is a change in *A*+'s circumstances which is such that, if *A*+ had been sensitive to it, it would undermine her being ideal.

Assume, then, that *A*+ may inhabit some situations featuring the circumstances of justice. If *A*+ were to lack a desire to cooperate, in many such situations, she would be punished as a free rider so that she would no longer be able to be instrumental rational – she might even become literally incapacitated, for example by being killed.[9] Clearly, that would undermine her ideal rationality. So *A*+ will only be *ROBUST*-ly disposed to be instrumentally rational if she has a desire to cooperate with other agents in situations where she may be punished. But her psychology should be *ROBUST*, and if she has the pro-cooperative desire, she will be able to maintain her ideal rationality. And as she may be punished in this way in all situations featuring the circumstances of justice – *ex hypothesi*, as she always may be overpowered – it follows that her psychology is *ROBUST* only if she has a desire to cooperate in *all* situations featuring the circumstances of justice.

To be clear, this is not to say that the ideal agent would be *more* instrumentally rational if she were to have a desire to cooperate. It is possible that she would be able to be instrumentally rational in at least some situations even without the desire to cooperate. Rather, with it, she is able be robustly instrumentally rational, which is needed for the manifestation of her capacities and dispositions to the extent which makes her ideal. It is *ROBUST* which does the magic trick here.

How should we characterize the desire that *A*+ must have to be ideal? For a start, the desire cannot allow her to cooperate when that seems instrumentally best and free ride when that seems instrumentally best. As the agents in the circumstances of justice are of roughly equal power, she would be able to be punished when attempting to trick others by free riding whenever she would be found out.

Would an *ideal* *A*+ always be potentially found out and punished? Yes. *A*+ cannot be smart enough to always be able to trick others. If we were to idealize her to that extent,

[9] Of course, situations where she gets killed are drastic, but they generate a very direct example of how the agent's dispositions and capacities may be destroyed by her lack of a cooperative desire, so they illustrate my point well.

It is, however, likely that something similar can happen to *A*+'s rationality in cases where she is not incapacitated or killed, such as when she is ostracized or imprisoned. In such cases, her instrumental rationality is *blocked* rather than removed, so it cannot be exercised or manifested. But her capacities must be fully functional or manifested, at least insofar as she acts, so blocked capacities also seem non-ideal.

to ensure a balance of power between agents in the circumstances of justice, we would have to idealize the other agents too. And we must do so, because that balance of power is an aspect of the circumstances of justice. This means that it will, in principle, always be possible for others to find out and punish even an ideal agent.

Furthermore, the desire to cooperate plausibly has to be final, and not merely instrumental, or else it would not be very robust. A merely instrumental desire to cooperate is unreliable and likely end up punished, since whether or not it is rational to enact often will be up for grabs, given the agents' other desires. For the same reason, the final desire must be strong enough for *A+* to act on it, or else she would not be able to be taken seriously by other agents.

With these considerations in mind, I take it that, to be fully ideal, *A+* must have a fairly strong final cooperative desire with a content which suggests that *A+* cooperates with others to satisfy *A+*'s other end-desire(s). Moreover, as an enabling condition for the successful and robust exercise of that desire, *A+*'s psychology must be *sensitive* to her situation. Sensitivity, in turn, imposes two conditions on her psychology. First, (*i*) *A+*'s desire to cooperate must be in one sense disjunctive; it recommends cooperation if others cooperate *or* acting on *A+*'s other (end-)desires if they do not. Second, (*ii*) *A+* must not (otherwise) have anti-social end-desires that would impede the exercise of the pro-cooperative desire.[10]

A+ is subject to these two extra sensitivity conditions because the desire to cooperate would not be possible to exercise successfully or robustly without them. First, cooperating with all agents, independently of their motives, would put the agent at risk of either being harmed by cooperation or a sucker's pay-off. On many occasions, this would undermine the rest of her psychology. But her psychology is supposed to be *ROBUST*. Hence, the desire to cooperate must be disjunctive.

Second, we could ask what would happen if *A+* were to cooperate on anti-social desires. By 'anti-social desires', I mean (final) desires for goals the satisfaction of which would significantly impede others' abilities to satisfy their own desires. For example, they might be desires to hurt others so that they cannot satisfy their other desires. If the desires to hurt others so that they cannot satisfy their other personal desires are satisfied, then those who are hurt cannot satisfy their own desires when cooperating. But then, the other agent(s) would not desire to cooperate with *A+*, given (*i*). It is obvious that, if such desires

[10] I say 'end-desires' – i.e. the desires to which the instrumental principle plausibly applies (cf. chapter 5, section 2) – here rather than 'desires' because it is possible that we still can have some weak anti-social desires that do not matter for our actions or reasons. Such desires need not be ruled out here.

are known among potential co-operators – which they will be by at least some co-operators in the circumstances of justice – aiming to cooperate on the desires will not have others wanting to cooperate with the agent who has them. So for the desire to cooperate not to be self-undermining, condition (*ii*) puts limits on *A*+'s other desires.

There are also some worries about how the desire to cooperate may look. First, one may wonder whether the argument for the desire to cooperate generalizes to other desires. Could not the ideal agent be punished for pretty much anything, and hence be rationally or functionally better off having the desires that others in the circumstances of justice would bully her into having? Second, might there be counterexamples where *A*+ is robustly instrumentally rational even if she decides to go it alone? And third, might it not be the case that it is enough for the ideal agent to cooperate with *some* cooperating agents? Maybe she is a Mafia boss popular with the other Mafia members. If *A*+ is that kind of Mafia boss, it is unclear why she would have to form a desire to cooperate with all other cooperative agents in the light of risking punishment for being a free rider in some situations. Maybe it is enough for her to cooperate with some other agents (such as the members of the Mafia)?[11]

All these worries may, however, be dismissed for the same reason – they are based on ignoring how *A*+ must have desires for what to do in the circumstances of justice. Regarding the first worry, as I mentioned in the discussion of (2) above, among the circumstances of justice is the fact that it usually is the case that agents are unable to perfectly satisfy their desires. We can interpret this as a given fact about how actual agents function: many desires are just brute, and not always easy to coordinate with others' desires. If so, the circumstances of justice allow that *A*+ would not be able to be bullied into submission by other ideal agents, for in the circumstances of justice, some desires are often just the agents' own. These are the types of desires that are hard to coordinate with others. It is for the sake of these desires that the desire to cooperate matters.

The second and third worries turn on another fundamental point about the circumstances of justice – namely, that I have stipulated that agents are roughly equally powerful, and hence can be overpowered by other agents. As far as the counterexample question goes, it therefore does not matter whether the ideal agent would be robust in *all* situations in or about which she must have desires for what to do. It is enough that she is robust in some situations for it to be the case that she needs the desire to cooperate in all of them, for any situation featuring the circumstances of justice is, *ex hypothesi*, such

[11] I want to thank Kirk Ludwig for raising the second and third worries here.

that the agent risks being overpowered by other agents. But agents always ward off the risk of being punished as free riders when they desire to cooperate.

Something similar can be said about the scope of *A*+'s desire to cooperate. In versions of the circumstances of justice where other agents are more or less close to *A*+, it is not enough to just cooperate with the agents who are close to *A*+ because she only risks being overpowered by these agents. This is because there, by stipulation, always are others who may overpower her. Real life Mafia bosses, famously, often end up in court.

It seems plausible to think, then, that a desire to cooperate must feature in *A*+'s idealized psychology. But how exactly does the desire to cooperate feature there? This takes us to a discussion of (5). Premise (5) says that if *A*+ must have a desire to cooperate with other cooperative agents, *A*+ possesses a desire to cooperate with other cooperative agents as a partially constitutive feature of her idealized psychology – and herself.

There are two options here. Either the desire to cooperate is ('just') an extra desire of hers, or the desire is part of *A*+'s instrumental rationality.[12] I prefer the former view. Extending IR_{HUMEAN} further would be very clunky. For then, instrumental rationality would require coherence between (end-)desires and means-beliefs *and* a particular desire, making it much less theoretically elegant than adding the desire to *A*+'s psychology.

Nevertheless, the desire should still be added to the ideal agent. It functions as an enabling condition for the robustness of her instrumental rationality, and this is because it allows her to take the best means to her ends in many different situations. So it is not at all *ad hoc* to attribute it to ideal agents. And because ideal agents are in part constituted by their psychologies, it follows that the desire is partially constitutive of the ideal agent. And then we get (C).

(3) From Cooperation to Morality

I have now argued that ideal agents must have a desire to cooperate in virtue of some constraints on what they must be like to explain reasons. This has not got us to morality yet, but we are getting closer. I shall now attempt to take the final step to get there. I shall not, however, speculate about what is moral, or respond to all possible objections one might have about how such norms might look. This dissertation is not about *what* is moral, but about sketching a way of explaining how some things may be.

[12] I guess it would also be possible for the desire to just be a necessary but not constitutive feature of the agent. But the ideal agent is constituted by her psychology, so that possibility is ruled out on my view.

Now, a moral norm, on my construal, is universally prescriptive and prescribes something that is conventionally recognizable as moral (cf. chapter 1, section 2; section 1 above). We can explain two such moral norms using the desire to cooperate. These lie at the heart of a more complex hybrid theory of moral reasons.

First, according to *DESIRES INTERNALISM*, reasons have their sources in the desires of ideal agents. With their desires to cooperate, idealized agents all have a reason-explaining desire to sensitively cooperate to satisfy their other (respective) end-desires.[13] So they all have a reason-explaining desire which suggests that they cooperate with other cooperating agents. Moreover, they also lack anti-social desires via the second condition, (*ii*), on the cooperative desire. Because reasons internalism says that the desires of an ideal agent explain our reasons, it follows that we all have a reason to cooperate to satisfy our other respective end-desires – except anti-social desires, which ideal agents are ruled out from having.

Since reasons are prescriptive, and all ideal agents have this pro-cooperative desire, guaranteeing that we all have the same reason to cooperate, it follows that we all have the same universally prescriptive reason. Anyone whose reasons can be explained by *DESIRES INTERNALISM* has a reason to cooperate with other cooperative agents.

Moreover, the reason is recognizably moral, for a fundamental pro-cooperative reason seems like the kind of thing we want to count as moral. We have a reason to cooperate in our social interactions so that we can act on our other desires (or reasons that they explain), which means that we have a reason to simultaneously *benefit* each other – we can act on reasons set by our respective desires – *recognize* each other's ends – since those are what we have reason to cooperate on – and *respect* each other as setting ends – for they are what we have reason to cooperate to satisfy. Moreover, since the desire to cooperate extends to all other cooperating agents, beneficence, recognition, and respect are mutual between all cooperating agents, so one may well argue that they are *fair*. These are familiar moral themes. Hence, the reason to cooperate explains a fundamental moral norm.[14]

[13] I have written: 'reason to cooperate to satisfy (…)'. Does this mean that agents have a reason to cooperate to *actually* satisfy each other's respective desires, or to cooperate in a way which *allows* them to satisfy their other respective desires? I think this is an issue in first-order ethics that my theory need not answer. Here it is enough to say that we have a reason to cooperate.

[14] There is a further interesting question here about the extent to which the reason would have to be dominant, i.e. count as stronger than the agents' other reasons – cf. how a 'dominant' desire counts for more than other desires in chapter 2, section 5, or Hare's property of overridingness, which I however mentioned that I am not committed to in chapter 1, footnote 7. But I take no stance on this issue since I do not want to commit myself to a theory of reason weights.

Second, there is also another moral norm that can be explained by the conditions that enable the cooperative desire. This norm stems from condition (*ii*). Condition (*ii*) rules out cooperating on anti-social desires, for it rules out anti-social end-desires. Given reasons internalism, it follows that it rules out some potential reasons. Moreover, it does so universally, since all ideal agents have it, and it is clearly recognizably moral, because it seems to explain a norm against anti-sociality in its own right. Hence, it explains a moral norm – but not a moral norm based on a reason; rather, it *rules out* some potential reasons.

To exemplify this, consider Bernard Williams' (1995) case of a husband who abuses his wife but lacks any motivation to stop (even after being idealized). Assume that the husband has a final desire to abuse her and lacks desires not to do so.[15] On Williams' view, the husband lacks an internal reason to stop because he cannot be so motivated. On my view, however, he would not have a reason to start in the first place, because his desire is anti-social. On any plausible interpretation of what 'abuse' is, the abuse limits his wife's abilities to satisfy her own end-desires when this desire is satisfied – perhaps out of physical pain, but more likely out of the psychological impact of such actions. This means that the idealized counterpart of the husband, whose desires give him reasons, will lack that desire. So he cannot have a reason to abuse his wife grounded in the desire to do so. Hence, I have been able to explain a second moral norm.

Can we say even more about the moral reasons we have on this theory, beyond the two moral norms? In section 1, I wrote that I was going to provide a hybrid theory of moral reasons. First, there is a dominant reason to cooperate. Second, what we cooperate on are end-desires, including moral ideals and practical interests that they generate, that do not conflict with criterion (*ii*). They are the desires (or reasons) we may cooperate to satisfy.

Because the view is structured like this, we can also explain an additional type of moral reasons. This type consists of the reasons that are necessary means to cooperate, and hence for acting on our fundamental reason, insofar as we are involved in social interactions. Given the reason-grounding desire to cooperate, a necessary means for cooperating is to cooperate to satisfy those other desires. Insofar as we are involved in cooperation-inducing situations, then, cooperating to satisfy them is a necessary means for living up to the central cooperative norm, for there is no other way to do it than through these desires (cf. Strandberg, 2019). These necessary means are *secondary moral*

[15] This is, of course, not a very realistic interpretation of all cases of abuse. But I am not after realism here, I am just after illustrating my point.

reasons. Secondary moral reasons are universally prescriptive because everyone has them (in the right situations) and are (at least usually) recognizably moral, but they are more contingent than other norms, for moral ideals and practical interests may vary.

One may, however, wonder: *how* are we to cooperate? We do not always have a clear reason to act in ways that *satisfy* each other's desires. But whether or not we fundamentally have reason to satisfy each other's desires or to allow each other to satisfy our own respective desires (cf. footnote 13 above), it is clearly the case that we do not always desire to have close interventions in our lives. If you and I both desire to compete and win in a game of chess, I do not have reason to let you win, and you do not have reason to let me win. Rather, we have reason to set up social structures, such as the rules of the game and competitions where we can play it, through which we can attempt to satisfy our respective competitive desires.[16]

Examples of such social coordination go beyond the case of games. They range from the very local to, in globalized times, the wholly global. If we desire to engage in close personal relationships, we have reason to lay down ground rules for how these should work. Such rules are not plausibly universally prescriptive – what happens in others' relationships does not need to be regulate my relationships. But other rules plausibly are. If we desire that future generations shall be able to live well (on this planet), we have reason to set up a social scheme that limits our respective climate impact.

What this proliferation of examples of social structures indicates is that what the moral norm of cooperation generates is a way to see our social interactions with others around us as components of a *system of mutual cooperation*. If we have that in place, we will have to organize ourselves in ways that suits our moral ideals and practical interests so that we can engage in practices that allow us to satisfy them.

However, I have not said anything about what our moral ideals or practical interests are. Combined with plausible assumptions about the content of our respective desires, lack of agreement, and so on, I suspect that something very much like a late-Rawlsian form of political liberalism (cf. Rawls, 1993) looks likely to come out of it, except applied to morality in general and only to a lesser extent to politics.[17] If that is right, my

[16] In fact, if games have constitutive aims, as I assumed when setting up the shmagency problem in chapter 3, section 1, we would not even be playing chess if we were to take such reasons into account. They are incompatible with the aim of chess, i.e. winning, or drawing if one cannot.

[17] At least, I believe our rules might end up looking like his after due reflection. It is probably right that we are caught up in ideological weaves that awkwardly shape our interests and ideals (cf. Haslanger, 2018; Walden, 2018). But we can approximate leaving them in virtue of reflecting on them well.

view can be given a moral contractualist interpretation. But I do not want to push this point. It is possible that we might desire other things than broadly Rawlsian societies.

(4) Objections and Replies

No doubt, many objections suggest themselves against the framework I just have presented. I shall conclude by discussing a number of them in the form of questions and answers.

Throughout this dissertation, you have mostly focused on individualist versions of constitutivism (and even postponed discussing social constitutivism early in chapter 2). Yet now you clearly appeal to facts about our social situation to generate moral reasons. So why did you dismiss the other views? Is it even coherent to 'go social' here? And can you explain universally prescriptive moral reasons if it is possible for agents, and ipso facto *ideal agents, to be in non-social situations?*

I shall tackle these questions in order. What about other, even more social, forms of constitutivism? First, when I mentioned social constitutivism in chapter 2, section 1, I wrote that social constitutivists seem to think that either some aspect of agency (e.g. rationality, personhood, or agency itself) is in part socially constituted, or they run more individualist constitutivist premises together with premises about the social nature of humanity to generate mixed views.

The problem with the former set of views, I suspect, is that they are based on too strong assumptions about some aspects of agency. As such, they probably suffer from versions of the problems of substantive assumptions and adequacy.[18]

Without discussing any particular views in depth, however, at this point in the dissertation, I can present a general argument that undermines them. For I have been able to defend a conception of agency without appealing to social premises. It follows from that defence that the stronger ontological assumptions of such views are undermined. At least the kinds of rationality or agency that can be used to explain norms do not *need* social dimensions, and there is no particular reason to assume that more would have to be added to do so. As such, there is some reason to suspect that nothing more than my premises is necessary to explain the normative phenomena I have attempted to explain.

[18] Not least, I suspect many such theories run into swamp man-style worries (cf. Davidson, 1987).

The second type of social constitutivism – that appeals to agency plus sociality, but where sociality still is distinct from agency – is closer to mine. Nevertheless, standard versions of such views tend not to be strong enough to generate universally prescriptive moral norms. This displays the second horn of the problem of substantive assumptions, which suggests that constitutivists may fail to explain moral norms if their assumptions are too weak. While it is debateable whether he is a constitutivist, we can exemplify this problem with David Gauthier's (1985) contractarian moral theory.

Gauthier starts off with a maximizing conception of rationality plus prisoner's dilemmas, and argues, based on them, that it is rational to have and act on dispositions to cooperate with others even on occasions where doing so does not maximize one's preference satisfaction. The conception of instrumental rationality he ends up with here is more than controversial (cf. Parfit, 1984, pt. I). But, on the other hand, if Gauthier were to stick with a more generic form of instrumental rationality, it is unclear why agents would find it rational to cooperate when they can free ride on cooperation. Again, in the circumstances of justice, it need not be individually maximizing to cooperate. So at least Gauthier-influenced forms of social constitutivism seem implausible, too.

But as I just have dismissed both types of social constitutivism, one may wonder whether it is coherent for me to go social. My answer is: yes, why not? I have appealed to our sociality – or the circumstances of justice – as a feature of our situation, which in turn is part of the explanation of our reasons. But I have dismissed other views for failing to respond to the problem of substantive assumptions, not categorically. So if my appeal to sociality is compatible with solving the problems that I have charged other social constitutivists with, there is no further worry here. And, indeed, I do not believe I am making an implausible assumption by saying that we are in the circumstances of justice. That point is part of my defence of premise (2) in the argument above. So appealing to sociality itself is not off the table.

What about the final worry? Am I able to explain universally prescriptive moral reasons if it turns out that some agents (or even shmagents) are not in the circumstances of justice? It does not seem like their idealized counterparts must have desires about what to do in the circumstances of justice. My answer is to introduce a degree of conditionality into my theory. *If* agents were to inhabit the circumstances of justice, then their idealized counterparts would have to have the right desires. This means that they would have the same reasons if they were there, whether or not they are. So they cannot step outside the moral community to avoid morality; their reasons would return were they to return, too. That is enough for universal prescriptivity.

A potentially awkward implication of this move is that it might seem like an agent who, in some sense, has transcended the circumstances of justice has no reason to go back there. We can stipulate that a non-ideal agent is in a position where she does not risk punishment for non-cooperation, and then her reasons to cooperate do not seem to apply to her in that situation. She may still have them on the condition that she enters the circumstances of justice, but she does not risk that, so there is no need to act on them.

However, even if an agent were to stand outside the circumstances of justice, it is unclear why the reason to cooperate would not be stronger than reasons not to do so. I am not committed to any particular theory of reasons weighing, but it seems very possible that a traditional theory of weighing based on desire strength, a moralized theory of reasons-weighing (Schroeder, 2007, ch. 7), or a theory that at least involves taking all agents' typical human desires into account (Manne, 2016), all will end up ranking moral reasons highly. The desire to cooperate would have to be fairly strong to matter (so it is weighty on the traditional view), it is clearly moral (so it is weighty on the moralized view), and all ideal agents have it (so it is weighty on the social view). Hence, cooperation will probably matter on any plausible theory of weighing.

It is still unclear whether you really have defended the best possible explanation of cooperative action. First, should you not understand cooperation as a kind of collective action, rather than as mutually disinterested means-ends cooperation that requires new desires? Second, might there not be some Gauthier-like interpretation of free-riding issues that does better than your view? And third, are there not evolutionary or social interpretations of how cooperation has arisen that need not involve rationality, but still are better explanations of the phenomenon?

I shall treat the three sub-questions in order. First, there is the question of collective action. It might be true that we should understand ordinary social behaviour in other ways than by positing individualistic subjects with Humean desires who need to cooperate on these. Perhaps we should assume that collective action and shared intentions are more social from the start.

I happily grant that there may be cases of collective action, perhaps via shared intentions, or perhaps via some other mechanism. It may even be that they ground their own specific obligations (Bratman, 2013) or provide their own solutions to prisoner's dilemma-like games (Bacharach, 2006). But that does not show that ordinary cooperation is not more business-like than shared, in the sense that it need not involve any kind of collective agency. It is plausible to think that we need to be able to cooperate with others on a much more basic, everyday level, than when we act collectively. Situations where we benefit from cooperation are ubiquitous, even when we lack special relationships with other co-operators.

The first sub-question, then, can be dismissed rather easily. The second one concerns Gauthier-style interpretations of the rationality of cooperation that, for all I have said, might be better than mine. As mentioned, Gauthier (1985; 2013) and followers (McClennen, 2004; Narveson, 2001) argue that considerations stemming from prisoner's dilemmas rationalize a kind of disposition to not always maximize individual utilities, but instead to cooperate with others. Should we not go with a take on cooperation like that one instead?

But it should be clear that I am not really discussing how to best adjust individual decision-making strategies when facing social dilemmas. Instead, inspired by Velleman (1997), I treat social interaction under the circumstances of justice as generating implications about how we, theoretically, best should understand the idealized agents whose desires give us reasons. Social interaction is not, in the way I discuss it, a problem which has impact on our principles of rationality.

The third sub-question concerns even more possible explanations of why we are cooperative, but this time from other disciplines. Attempts to explain action or behaviour that is moral, in my sense, or 'pro-social' or 'cooperative', as it is sometimes called, have become commonplace in many disciplines. There are attempts to do so all over the social and mind sciences, and even in evolutionary biology. Among other things, it is often argued that we tend to change the games we play from one-shot prisoner's dilemmas to others, where cooperative strategies are better (Ostrom, 1990; North, 1991), or that we have evolved to sympathise with each other because of the benefits of pro-sociality (de Waal, 2006; Joyce, 2006). Is my action-theoretical argument at all helpful here?

One might think that a debunking argument against me might come from these more scientific explanations. If altruism, cooperation, and similar phenomena are better explained by various scientific theories than by my pro-cooperative desire, then one might want to do away with my arguments. One may, for example, say that it is more plausible to think that such behaviour stems from evolved impulses (Joyce, 2006). Moreover, versions of this argument might be thought to make sense even if they are supposed to vindicate morality of some sort (cf. de Waal, 2006). For if so, one would have a story about morality which is independent of the workings of pro-cooperative ideal desires.

My answer has several parts. Regarding the evolutionary hypotheses, there are many *prima facie* moral phenomena that my framework can accommodate. There is no tension between arguing that morality has evolved through moral emotions and positing moral ideals based on individual reasons for action. But moral emotions do not, by themselves, give us an explanation of the normativity of morality. To explain that, rationalized desires still seem helpful, even though not all moral phenomena originally may have been caused by them.

Second, if one were to opt for some more scientific explanation of cooperation, it seems like my arguments in the theory of reasons provide an interesting causal explanatory hypothesis in their own right. Various forms of cooperation often feature in explanations of morality, and enhancing group coordination for the purpose of satisfying various desires is often recognized as a collective aim. Even Railton's feedback mechanism (cf. chapter 6, section 2) provides a story about how it might be the case that we are responsive to our (cooperation-oriented) reasons whether or not we think we are. Hence, it does not strike me as at all implausible to think that much of the cooperative behaviour that we see instantiated among humans – and maybe even among other higher animals – depends on their adjusting their desires to participate in shared cooperation. Obviously, this theory is untested. But it would not be *ad hoc*, due to the theoretical arguments in favour of it here. So why not give it an empirical shot?

Fair enough, this might suffice as a general view. But there is still a worry that remains given your constitutivist framework. It seems like the shmagency objection might reappear.

In fact, there are several distinct shmagency worries here. You presented two desiderata *for any plausible reply to the shmagency objection in the conclusion of chapter 3. First, a constitutivist aiming to explain reasons should be able to explain reasons for (at least some) shmagents, like the Martians and the Saturnians – and we may add, plausibly, also for* agents *who are unable to perform paradigmatic actions, if there are such agents. Second, such an explanation should avoid the problem of underdetermination. How do you handle these* desiderata*?*

First, remember my general theoretical picture. I defend a two-tiered form of constitutivism, where the first tier features instrumental rationality as a constitutive feature of paradigmatic action and agency, and the second one explains reasons in terms of the idealized responses of paradigmatic agents. And I have also argued that the instrumental principle is plight-inescapable (chapter 5, section 6), which explains its normativity and is part of the explanation of its dialectical inescapability. Given the considerations in chapter 3, section 2, it follows that we cannot shirk from rationality – but there may be non-paradigmatic agents or shmagents who lack it.

How do I handle the *desiderata* I set up for constitutivist accounts of normative phenomena in the conclusion of chapter 3? The first one said that a theory of reasons should explain reasons for shmagents, along with the reasons of agents. Now, I have not tried to explain rationality for shmagents, but instrumental rationality is not or does not ground a reason for action – it is another kind of normative phenomenon. It does not seem strange that there can be shmagents, like the Martians or Saturnians, without it. They lack it *ex hypothesi*.

Is this too simple? After all, I did argue that they seem to have normative reasons since they can engage in ordinary normative practices. So why should they not have some sort of principle of rationality built into their psychology, for the same reason? The answer comes from the role I have argued that instrumental rationality plays in an agent's psychology. It connects beliefs and desires in a way that shows how agents stand behind their actions, but it is possible to perform bad – or non-fundamental – actions (cf. chapter 5) that do not feature it. As such actions are possible, it must also be possible to be such that one only can perform them. It therefore follows that there are 'shmagents' relative to *PARADIGMATIC AGENCY*$_{IR}$ who only can do that, but who are unable to perform paradigmatic actions. The Martians are Saturnians are plausibly like that. But they cannot have a principle or capacity for instrumental rationality, for then they would *ipso facto* be agents. So instrumental rationality has a special categorical role that reasons lack. It differentiates agents from shmagents.

What about the second *desideratum*, i.e. showing why the norm I have explained is not underdetermined? Paradigmatic actions involve being fully instrumentally rational, and failing to be instrumentally rational is to act in a way which counts as a non-fundamental kind of action (cf. chapter 5, section 7). So the norm of instrumental rationality is not underdetermined; one may only either succeed or fail to act on it, where the actions count as different kinds of action. Why not perform a non-paradigmatic action? Because paradigmatic actions are plight inescapable (cf. chapter 5, section 6).

The second tier of my view is more complicated here. It involves a form of partial constitutivism (cf. chapter 3, section 5; 6): reasons are explained by the responses of idealized agents, where the idealization is understood (in part) in terms of rationality. This means that we need not be normatively perfect ourselves to be able to explain reasons. But, to start with the second *desideratum*, why should our reasons be explained by the desires of counterparts that desire to cooperate, as opposed to less idealized counterparts that do not? Assume even that our reasons stem from ideal agents with the desire to cooperate. If so, are not shmeasons enough for us? Why does it matter that we act on reasons with their sources in the desires of fully ideal counterparts rather than ones that are slightly less than ideal? This is the shmagency objection I pushed against partial constitutivism in chapter 3.

My answer lies in the constraints set out in the argument itself. If *A*+ is suitably idealized, and hence able to explain reasons, *A*+ must be *CLOSE* and *ROBUST*. Much argumentative work is done by the idealization that explains these properties; idealization is what explains many of the features of reasons I discussed in chapter 6, and getting rid of these properties would impair our explanation of our reasons (cf. section 2 above). A less ideal idealized agent, lacking *CLOSE* and *ROBUST*, cannot plausibly explain our reasons.

Moreover, the desire to cooperate is *legitimized* by *A*+'s instrumental rationality. Hence, it cannot be ignored as reason-giving by agents in this world. It is therefore normative. Might some other desires be legitimized instead? That possibility was, after all, the way in which this shmagency worry arose for Smith. But the answer is 'no'. The desire to cooperate cannot be legitimized by different amounts or kinds of instrumental rationality on part of the agent. This point stems from the disjunctive nature of action. Ideal agents are fully functioning paradigmatic agents, meaning they have all the properties constitutive of paradigmatic action, and manifest them fully insofar as they act. Other actions belong to other kinds of action (cf. chapter 5, section 7), but the ideal agent does not perform those kinds of actions. Hence, they have the type of instrumental rationality that legitimizes their desires, including this desire, rather than some other set of desires.

In general, my point here indicates an important difference between my view and the views of philosophers like Smith, Velleman, or Katsafanas, who hold that the constitutive feature that explains norms comes in degrees. They explain reasons in terms constitutions that can be more or less fully instantiated. But here, we can only succeed or fail to be instrumentally rational; there are no degrees of success. There is no legitimizing alternative to explaining reasons in terms of the desires of a fully instrumentally rational counterpart. Hence, my explanation of the normativity of reasons makes good on the second *desideratum* I presented for a good answer to the shmagency objection. The disjunctive view avoids the problem of underdetermination.

Is the solution too quick? One problem for my view is that it still seems possible for an ideal agent to be less than fully *idealized*. Why go with the reasons provided by fully idealized desires rather than other desires, especially since both sets can be compatible with rationality? The answer is that I am not really committed to full idealization – that could, in principle, involve just about anything. I am committed to idealization insofar as it explains reasons, and a theory which is not idealized in these ways would be very implausible. Hence, the desires legitimized by instrumental rationality and possessed by ideal agents (as I understand them) can explain reasons, whereas the desires of other rational agents cannot.

One may ask 'So what?'. It is one thing to argue that agents who are instrumentally rational must be robustly instrumentally rational when ideal, and hence that we can explain their reasons. But what about the reasons of creatures who lack the capacity to be rational? This takes us to the other *desideratum* I presented for a solution to shmagency problem. The Martians or Saturnians seem to have reasons, even though they are not agents. Moreover, we may simultaneously want to think about the reasons of agents who cannot take means to ends in any way. If they exist, do they not have reasons?

It is unclear where shmagency ends and pure objecthood begins. I shall assume that any shmagents who can have reasons (or something reasons-like) will be able to have representations of means and ends, and that they tend to behave (i.e. almost-act) based on combinations of these. Hence, both the Martians (with beliefs and desires, but not *IR*) and Saturnians (with besires) are shmagents of the relevant kind.

I hypothesize that we can explain the reasons of means- and ends-representing shmagents in the same way as the reasons of agents, i.e. by appealing to the desires of idealized agents, assuming their desires and beliefs are given the same content as the means- and ends-representing states the shmagents have. For remember my argument for *DESIRES INTERNALISM* in chapter 6 – idealized desire-based reasons explain all we need to explain to capture reasons. And we know from chapter 3 that the reasons the Martians and Saturnians have seem suspiciously much like human reasons. If so, *DESIRES INTERNALISM* would generate a suitably unified explanation of reasons.[19]

[19] Another way to explain reasons in a unified manner would be by appealing to idealized shmagents. I do not think this approach works, however, because reasons must be normative, and the shmagents need not have access to instrumental rationality, which is what legitimizes their ends.

Indeed, *mutatis mutandis*, I suspect it can do so for *any* creatures that have representations of means and ends. The explanations of most phenomena explained there work straightforwardly in the same way. For example, the reasons bear relations to the agents' idealized ends, capturing a property such as *RELATIONAL CHARACTER*, even though these ends are represented by desires among the idealized agents, instead of being represented by (e.g.) the besires that some shmagents have. But nothing seems to differ between these sets of mental states, extensionally speaking. And similar stories can be told about all the other features of reasons that I tried to explain – I leave working out the exact details here to the reader.

There is one final complication, however. What about agents who only are able to perform non-means-ends actions? It would be easy to suggest that they cannot have reasons for actions just because they are unable to perform means-ends actions. There is something to be said in favour of that view. They seem unable to perform the paradigmatic actions that paradigmatic agents can perform, so their psychologies are importantly different. On the other hand, *if* there are non-means-ends actions, it is not implausible to think that there may be reasons for them.

In principle, however, there is nothing about *DESIRES INTERNALISM* that rules it out from explaining the reasons of agents who only can perform non-means-ends actions. Part of the hypothesis is that it explains all reasons for action for Humean agents, including their reasons for non-means-ends actions (if there are such actions). But then it does not seem like it would not be able to explain the reasons for non-means-ends actions that non-Humean agents have too, even though their abilities to act on them may be fairly limited. So *DESIRES INTERNALISM* remains a living option here.

Still, I would also be happy to restrict my constitutivism to a theory about the reasons of paradigmatic agents (and shmagents who are much like them). Even though one arguably should explain the reasons of shmagents who are fairly similar to us, why one would have to explain the reasons of wildly different creatures using the same machinery is unclear. So whether *DESIRES INTERNALISM* should be extended to cover such creatures is a question for the future.

(5) Conclusion

In this chapter, I started out by introducing some terminology in section 1. Then, in section 2, I presented the argument from idealization. I argued that idealized agents must be *CLOSE* and have psychologies which are *ROBUST*. From those properties, and some other background assumptions, it follows that they are partially constituted by final desires to cooperate. In section 3, I attempted to explain at least two fundamental moral norms based on the properties of that desire – one norm suggesting cooperation, and one limiting the desires we can have to non-anti-social ones – as well as less central norms that depend on the agents' interaction with other agents. Doing so, I defended a hybrid theory of moral reasons. Finally, in section 4, I replied to some objections. That concludes the last substantive chapter of this dissertation.

8. The Case for Constitutivism

The substantive discussion of this dissertation is now over. In chapters 2 and 3, I argued that the leading older versions of *FORMAL CONSTITUTIVISM$_{PM}$* fail in virtue of the problems of substantive assumptions, adequacy, and shmagency. In chapters 4-7, I have instead tried to develop an alternative, Humean, constitutivism.

Here, I aim to sum up the case for my view. In the first two sections, I show the benefits of constitutivism. I start in section 1 by showing how my form of constitutivism solves sceptical problems, and in section 2, I consider the quality of its explanation of action and agency. I then summarize how my form of constitutivism improves on older forms by solving problems for them. In section 3, I discuss how it improves on Korsgaard's, Velleman's, Katsafanas' and Smith's versions of constitutivism by solving the three most pertinent problems they suffer from. And in section 4, I provide general solutions to the remaining standard problems for constitutivism. In section 5, I conclude this chapter (and dissertation).

(1) Sceptical Worries in Metaethics

The first of the positive arguments for constitutivism comes from its ability to avoid sceptical worries. I shall argue that constitutivism can solve the sceptical worries I consider to be the most pertinent, i.e. the metaphysical and epistemic queerness-based worries I originally presented in chapter 1, section 3.

Metaphysical Queerness

The most important sceptical worry about morality, as I see it, is the well-known difficulty of explaining why moral norms are universally prescriptive – or, in other words, explaining why they have normative force or 'oomph' for all. But before I can get to an explanation of that, there is another problem. Some error theorists, such as Olson (2014, ch. 5-6), think that the irreducibility of normativity is a deeper problem still (cf. chapter 1, section 3). How do I handle these issues?

My solutions to both problems should be apparent from how I have developed my view in chapters 5-7. I have argued that rationality is structurally prescriptive (chapter 5, section 5; 6), that reasons are prescriptive because they are grounded in rational desires, where rationalization legitimizes them (chapter 6, section 2), and – most importantly – that moral norms are universally prescriptive because they have force for all (chapter 7, section 3). The positive norm of cooperation is universally prescriptive because it is a reason for all, the negative norm of non-anti-sociality is universally prescriptive since it is a condition limiting all ideal agents' desire sets, and secondary moral norms are necessary means for living up to the norms just mentioned. Hence, the strategy starts by establishing an agent-relative norm (of instrumental rationality) which then is used to explain further norms, including universally prescriptive ones.

How reductive is this view? No appeal has been made to normatively irreducible properties in this explanation. But the theory need not be reductive across the board. It is compatible with thinking that psychological states may be irreducibly non-physical – and possibly even irreducibly normative (cf. section 3 below).

Nevertheless, even if mental states are not reducible – and I say that even though I suspect they are – my theory remains *normatively* reductive all the way down. For even though mental states may be irreducibly normative, their normative *force* comes from how they are plight inescapable and otherwise entangled with our agency (cf. chapter 5, section 6). So even if they are irreducibly normative, the sense in which they are normative is no weirder than the sense in which grammatical norms are. They may apply to people without having normative force for them.

Moreover, the view I have defended is normatively reductive in an especially fruitful way. Because it involves a hybrid theory of moral reasons, it captures practical interests and moral ideals along with more central moral norms, and it shows how they, together, serve to generate secondary moral norms. This means that while other views also might explain normativity without queer assumptions, my view is likely to be a serious contender even in contrast with those theories.

Epistemic Queerness

The second major sceptical worry for morality is epistemic queerness. I claimed in chapter 1, section 3, that I would discuss three forms of epistemic challenges. First, there are worries posed by the method of reflective equilibrium, indicating that it would be nice to show how constitutivism is epistemically central. The second two are the problems of moral disagreement and 'cosmic coincidence', which are special problems for moral knowledge. I discuss them in this order.

First a *caveat*, however. To repeat, I am not committed to coherentism. Even though I will say some things about reflective equilibria, I do not mean to do so *qua* coherentist, but rather because it is an important general philosophical methodology. Its problems generalize even beyond coherentism.

I emphasized three general problems for the method of reflective equilibrium in chapter 1, section 3. These are the garbage-in/garbage-out problem, the problem of truth-conduciveness, and the problem of multiple reflective equilibria. However, my type of constitutivism shows how moral norms depend on (idealized) mental states. This argument provides the right kind of explanatory link between our psychologies and the norms to show how they hold and have force for us. Hence, the argument from idealization – and its moral consequences – show how moral reflection has a starting point which is not garbage. It is a central anchoring point in our moral thought.

Moreover, *qua* transcendental argument, the argument from idealization is also truth-conducive (if its premises are correct). As I believe its premises – or something like them – are correct, it follows that the problem of truth-conduciveness has an immediate solution too.

Nevertheless, it might be thought that my moral picture cannot solve the problem of multiple equilibria. It is very likely that there are many different desire-based sets of secondary moral norms. Hence, we might end up thinking about these in awkwardly divergent ways.

There is a sense, however, in which the coordinating agreements that we will reach, or at least would reach on reflection, are those that hold for us. This means that it is quite possible that several different kinds of norms *possibly* may hold for people, according to my theory. But because secondary moral norms depend on what people agree on, there is a sense in which their agreement *ipso facto* settles what the facts are about what the secondary moral norms are.

True, to gain knowledge about what people would settle on, one would have to take facts about other agents' idealized motivational sets – or reasons – into account to think about the norms that will hold, since these have impact on which agreements people will reach. But that is not a principled worry. Facts about desires or reasons are knowable.

Might it be *too hard* to come to know what the agreements are (or would be)? No, constitutivism becomes normatively central in a second way here. Because it shows that some normative judgements are transcendentally justified, these also become *psychologically* central. For insofar as they are accepted, they are likely to generate further moral judgements (and intuitions, emotions, etc.) in line with them among those who accept them.[1] This means that the kind of judgements people make will tend to become socially accepted and that we will think in terms of them, over time. This should inspire some hope for thinking that we can come to know the agreements.

So much for the centrality of constitutivism as a way to ameliorate problems for the method of reflective equilibrium, then. What about the more particular epistemic challenges, i.e. those of moral disagreement and cosmic coincidence?

The question of disagreement has already received a partial answer. I argued that constitutivism is central both for providing a transcendental argument for some moral norms, and for making agents who accept it moralize in line with it. This means that agents have no reason to disagree, fundamentally, on some central moral norms.

Even so, some moral disagreements will no doubt remain. Moral disagreements often seem especially pertinent, after all. But the structure of the central cooperative norms provide us with reason to think that disagreements between constitutivist agents cannot be fundamentally deep. They have reason to cooperate, and, insofar as they are ideal, cannot have reason-giving anti-social desires. This means that they have a strong reason to find some sort of cooperative agreement – even to the extent that some of them might have to change their respective desires if those are incompatible with their core reason to cooperate, as cooperating is what they have reason to do. Accordingly, there is strong normative pressure for constitutivist agents to reach cooperative agreements. And as there is no principled worry about knowing people's desires or reasons, their agreements are in principle knowable, too.

The challenge of cosmic coincidence is also easy to solve. My kind of constitutivism does not deal in moral facts beyond those that depend on idealized mental states. There is nothing strange about coming to know them – they are not epistemically distinct from anything else. Knowing them only requires our standard epistemic abilities.[2]

However, Barkhausen (2017) has tried to generalize the cosmic coincidence worry to apply not just to the metanormative non-naturalism it is usually addressed to, but also to forms of naturalism. His worry is that we only can be shown to have reliable moral beliefs if we accept presently held moral opinions as justified, leading to unjustifiable first-order moral consequences. Current moral opinions are likely to be mistaken, he thinks.

[1] I presume it is uncontroversial that accepted moral values tend to influence people's psychologies.

[2] In fact, the psychological centrality I have argued that constitutivist transcendental arguments have is likely to rein in our intuitions and incline them towards fitting cooperative norms, too.

If push came to shove, I could accept that point. My argument for constitutivism does not rely on assuming that other moral norms are justified. But there is also a deeper solution, stemming from the fact that, desires to cooperate aside, (ideal) agents will cooperate on their end-desires, where these cannot be anti-social desires. Those desires are not likely to be morally hideous from the standpoint of our present moral views. They are innocent because they are not anti-social, and they are likely to both cause and be caused by otherwise accepted moral opinions. This means that moral norms that capture them are unlikely to be hideous – or, indeed, contemporary:

(2) Explanatory Benefits

Beyond its abilities to solve problems in metaethics, the second main positive argument for constitutivism is that it might be based on a good theory in the philosophy of action. If normative requirements indeed can be explained by some aspects of agency, and the latter are defended, the requirements follow. Are my action-theoretical premises good enough for that?

Sort of. I want to say 'yes', but the 'yes' has to be very tentative. If the arguments for the Humean theory of agency and for how an idealized version of it explains reasons are sound and valid, constitutivism would follow. But are they sound and valid? Everything positive I have said remains extremely controversial, not least just because of what philosophy is like, but also in the light of the objections I have tried to deal with in the appendices and chapter 6.

Hence, it seems prudent to say that the 'yes' in virtue of which I endorse my conclusions comes with a qualification. Because of how controversial my arguments remain, I only think of the views I have defended as tentatively and inconclusively justified. Until they can be much more comprehensively defended, I only view them as living theoretical options – albeit tentatively justified ones – which, I soon shall argue, may be helpfully adapted in several metaethical frameworks.

(3) The Key Problems

The three main problems I raised for standard constitutivist views in chapter 1, section 4, were the problem of substantive assumptions, the problem of adequacy, and the agency-shmagency problem. With my positive case for constitutivism now presented, I return to them in the same order. Insofar as my view can solve these major problems for constitutivism, it is preferable to forms of constitutivism that cannot.

The Problem of Substantive Assumptions

The problem of substantive assumptions is a dilemma. It seems like the more substantive assumptions constitutivists make to explain normative phenomena, the easier it is to deny something that is assumed. But the less substantive assumptions they make, the harder it seems to get any substantive norms out of constitutivism.

By and large, I have aimed to go for the second horn. I have wanted to make weaker assumptions than other constitutivists, and hence I have appealed to a Humean view to supersede the flaws of previous accounts. In chapter 2, I argued that Korsgaard makes too strong assumptions about universalized maxims, Velleman about the relation between self-knowledge and control in action, Katsafanas about drives, and Smith about which desires to add to ideal agents. But my Humean framework is not committed to their assumptions (cf. chapter 4, section 4).

One might, however, think that I have made some strong assumptions myself. In particular, this is so regarding the components of *HTM* and idealization. My response is that my assumptions still are weaker than those that others have made – so I have superseded their problematic assumptions. And I have defended my assumptions at length. Hence, to the extent that my assumptions may be construed as substantive, I have been able to defend them in spite of the challenge stemming from the first horn of the dilemma.

The Problem of Adequacy

The second major problem for constitutivism is the problem of adequacy, viz. the problem of actually being right (or at least theoretically adequate). This problem is intimately tied up with the problem of substantive assumptions. In chapter 2, I argued that leading constitutivist views suffer from the latter problem *because* their defences of their assumptions are inadequate (in the respects I just mentioned when discussing the last problem). By contrast, I believe my view is theoretically adequate in the light of my defence of it.

However, as I also have indicated, that defence is tentative and inconclusive. There is more work to do to defend my view comprehensively. Nevertheless, I have defended my key assumptions, and hence take them to be justified – if ever so tentatively.

The Shmagency Objection

In chapter 3, I charged standard constitutivist views with not being able to respond to the shmagency objection. The standard reply from self-defeatingness does not work because there can be sophisticated shmagents, and partial constitutivist responses suffer from a problem of underdetermination.

Yet I have presented a two-tiered form of constitutivism that seems to run into versions of this problem in its own ways. First, I have argued that paradigmatic action, and hence paradigmatic agency, is constituted by beliefs, desires, and instrumental rationality. How does that view handle the *desiderata* that I argued that any solution to the shmagency objection must handle? Second, the second tier of my view consists of a kind of partial constitutivism according to which reasons for action are grounded in idealized desires. How does *that* view handle the *desiderata*?

In chapter 7, section 4, I explained how it handles them. The first *desideratum* says that if one is after explaining reasons, one must be able to show how sophisticated shmagents have them. But instrumental rationality is not supposed to be or ground a reason to be rational, so that *desideratum* does not apply here. True, my sophisticated shmagents are not rational. But nothing of normative interest follows from that.

The second *desideratum* says that the normative phenomenon some form of constitutivism explains should not be underdetermined. Here, instrumental rationality is constitutive of paradigmatic actions – and actions that do not belong to it belong to non-fundamental kinds of action. It follows that there is no alternative interpretation of it or any alternative principle one may follow instead, for then one is no longer acting paradigmatically. And why could one not act non-paradigmatically? Because paradigmatic action is plight inescapable.

I have also presented solutions to the *desiderata* for the second tier of my view. Because instrumental rationality does not come in degrees, the reasons it explains are not underdetermined. Hence, it makes good on the second *desideratum*. And I also argued that this theory of reasons, suitably generalized, seems likely to be able to explain the reasons of all relevant shmagents (and possibly even the reasons of agents who cannot perform means-ends actions), hence making good on the first *desideratum*. So my view does better than the constitutivist views in the literature when it comes to answering the shmagency objection.

(4) General Objections

In chapter 1, section 4, I also raised the problems of bad action, alienation, and contingency, as well as the is/ought and metaethical problems, for constitutivism. When it comes to all these problems, however, my responses may be used to develop other forms of constitutivism than mine as well. So my responses here do not constitute arguments in favour of my preferred view over other forms of constitutivism; rather, they are free for the taking for all constitutivists.

The Problem of Bad Action

Generally formulated, the problem of bad action is the problem of explaining how one can act (or otherwise instantiate some aspect of agency) in a bad way if action (or some aspect of agency) is normatively constituted. If all actions, for example, involve following *CI*, how do we account for actions that fail to live up to it?

In my case, the problem turns up for the first tier of my theoretical package. If instrumental rationality is constitutive of agency, what do I make of irrational agency? My answer is presented in chapter 5, section 7. It is disjunctivist, because it takes successful versions of belief/desire-motivated action to be the fundamental kind of action, whereas failed action belongs to some other kind(s).

This solution is slightly different from most traditional solutions to the problem, including the standard so-called 'threshold solution', according to which all actions must pass a threshold of minimal success to count as actions (Lindeman, 2017). But on the disjunctivist view, there can be actions fail to live up to the constitutive features of agency in any way, which threshold theorists cannot say. This case justifies the disjunctivist view.

I do not see why other constitutivists could not adopt this solution, however. I think it would improve all the leading constitutivist views. In particular, it improves on Smith-style ideal response explanations of reasons. This is because it generates a solution to the underdetermination version of the shmagency objection – which has been especially helpful for me. Because rationality does not come in degrees but rather in kinds, it saves my version of a similar view from suffering from the shmagency worry that I have presented for his version of constitutivism.

The Problem of Alienation

The problem of alienation is the problem of explaining why one could not half-heartedly, grudgingly, be committed to constitutivist normative phenomena. If that is possible, their normativity does not seem to be captured merely by the fact that one is (inescapably) committed to them. The constitutive features of agency might be some sort of delusion; we have inescapable commitments, but they need not reflect a deeper normative reality. Rather, while agency may be inescapable in some sense or another, what we are committed to need still not be what we fundamentally ought to be concerned with.

This objection has sometimes gone under the name of 'half-heartedness', and then been presented as a version of the shmagency objection (Enoch, 2006; cf. Tenenbaum, 2019).[3] But it is distinct from the shmagency objection, for I take the shmagency objection to concern whether agency is in some sense inescapable, whereas the inescapability of agency is conceded in the alienation worry. Hence, I have opted to treat half-heartedness or alienated participation in agency separately from the shmagency objection.

It seems unfair to label normative phenomena we are committed to in virtue of our constitutions 'delusional', however. Constitutivist norms are grounded in properties that are constitutive of features of aspect of agency. These are facts of human life. It would be one thing to be committed to certain beliefs just in virtue of believing. These could be false. But there is no sense in which instrumental rationality, as in chapter 5, or desire sets constitutive of ideal agency, as in chapter 7, can be false. They are just there.

Then again, there might also seem to be a sense in which normativity goes deeper than instrumental rationality or desires if one has a background theory of what might be normatively right. The constitutive features of agency need not be related to what it might imply. Hence, one might still worry that constitutivism does not reach down into actual normative reality.

[3] More specifically, Enoch (2006) wrote that we could pursue whatever is constitutive of agency half-heartedly, not attempting to act on it well, in virtue of other ends we may have. Tenenbaum (2019) has replied that while constitutivists can reconstruct such cases as showing that the agent does not really seem to be aiming at what is constitutive of agency, what is constitutive of agency can still come into conflict with other aims the agent has, and so the agent should be able to give precedence to her other aims. What interpretation of alienation one opts for, or whether these ideas fundamentally differ, is irrelevant here.

But presuming that there might be a true theory of what normativity is in the background, independently of agential commitments, ignores the dialectical context in which constitutivism appears as a theoretical option. In particular, constitutivism acquires a lot of its strength from its ability to respond to sceptical worries in metaethics that challenge views according to which normativity is supposed to be reflective of some deeper, underlying reality. This means that there is no presumption to be had in favour of such a reality.

Rather, constitutivists are more plausibly construed as trying to explain how properties of constitutive features of agency might play fundamental normative roles on the assumption that there are significant problems with views that take there to be a deeper normative reality. And then, as I shall argue below when discussing the metaethical problem, the features of constitutivist frameworks might themselves be given various metaethical interpretations.

The Problem of Contingency

Another problem that besets constitutivists is the problem of contingency. Ideally, one might think, at least moral norms apply or have force over several, if not all, possible worlds. They are normative with a kind of necessity, and so should remain the same even in different circumstances (cf. Fine, 2002; Horgan and Timmons, 1991). But according to most forms of constitutivism, this does not seem plausible. Agents – or at least shmagents – with other constitutions than ours are possible, and hence constitutivism does not seem to be generally applicable.

Still, constitutivists should be able to bite this bullet. Inspired by Williams (1985, ch. 9), we can endorse a 'relativism of modal distance', where it is enough to be able to explain (at least moral) norms for agents (of some particular kind), and possibly shmagents who are similar to them, whereas other agents (or shmagents) that are different enough well may be subject to other norms. I have appealed to a version of this kind of relativism of modal distance myself when I discussed the shmagency objection. It is plausible that some version of my theory of reasons can capture reasons for shmagents who can engage in means-ends behaviour, but possibly not the reasons of agents or shmagents who cannot engage in means-ends behaviour.

Does this mean that constitutivists cannot capture morality at all? What about the appearance of necessity that morality seems to have? The alleged datum that moral norms appear necessary can be given a constitutivist explanation that does not appeal to their logical, metaphysical, or normative necessity. We might have this intuition because the agency that explains moral norms seems inescapable *for us*, which shows why morality seems to hold with some kind of necessity. It would hold for us regardless of what situation we are in insofar as we remain agents of the kind we are, but not necessarily for all possible agents or shmagents (cf. Street, 2012; Lavin, 2017).

Moreover, my response to the shmagency objection suggests another way in which constitutivists can explain how moral norms seem to hold for any creatures like us. One might hold that any creatures like us should have their norms explained in the same way as ours, whether or not they are agents or shmagents. In practice, this means that all creatures who are able to take means to ends also should have the same moral reasons – one would have to be unable to take the means to one's ends to not have similar moral reasons. This means that the difference between us and the kind of creatures for whom morality might look different is rather significant.

Is there a risk that there might be some creatures who can perform means-ends actions but who lack the relevant moral reasons? In chapter 7, section 4, I argued that it is likely that all humans will inhabit some circumstances of justice, and that even a creature outside the circumstances of justice well may have a weighty reason to cooperate. This response also provides the key for thinking that the relevant reasons are universal (in at least the sense that they hold for all creatures who can act or behave in ways where they take means to ends). For while a creature that stands outside the circumstances of justice is conceivable, it is very likely that even the ideal agent of such a creature will have desires about what to do in at least some situations that feature the circumstances. If that is right, the rest of the argument from idealization can be re-run, yielding these creatures a reason to cooperate.

The Is/Ought Problem

Do constitutivists straddle Hume's famous is/ought-distinction by attempting to explain norms reductively, in terms of non-normative properties (Hume, 1978, pp. 469-470)? The nature of this divide is contested (cf. Baillie, 2000, ch. 5). There are two main versions. Either the divide is metaphysical, concerning the way in which descriptive facts may (or may not) explain normative ones. Or, alternatively, it is (broadly) semantic, in which case it concerns the possibility of making normative inferences.

Constitutivists should be fine with the former worry. They need not take a stance on the issue. For they can happily hold that the psychologies of ideal agents – or whatever else explains constitutive norms – are not reducible. Perhaps it is impossible to talk about human psychologies without appealing to norms, such as the standards of success set by beliefs aiming at truth or desires aiming at satisfaction, let alone the instrumental principle. If so, constitutivists have been making moves inside the normative sphere all along.

However, they may well also think that psychologies are so reducible (cf. Smith, 2017). I am inclined to think so myself. If this turns out to be right, then so much the worse for the metaphysical is/ought-divide.[4] But, as I mentioned when discussing metaphysical queerness worries in section 1 above, whether the mental is wholly reducible is a question which is far beyond the scope of this dissertation.

What about the latter, inferential, worry? The problem is that we might seem to be inferring things like '*A* ought to φ' from some descriptive sentences (or whatever else one infers something from) about the nature of agency. But constitutivists can make largely the same moves here as in the metaphysical case. That in virtue of which the normativity of norms is explained might actually be normative, too, and hence picked out by normative sentences. For example, I have argued that reasons can be explained in terms of idealized desires, and maybe one should interpret those desires as desires one *should* have, in some sense of 'should'. Then those desires also plausibly explain what one should do – *that* can follow logically, at least given other suitable premises.

But then, that normative feature may or may not be normatively irreducible. With these options in mind, constitutivists can argue either that the inference was descriptive all along (if some reductive view is true), even though that is not obvious on the surface, or that the inference from some descriptive facts about agency in fact can be coupled with some normative premise to get further normative conclusions. So there is no inferential problem here either.

The Metaethical Problem

One recurring theme in discussions of constitutivism is that it is unclear where it should be located in the metaethical literature. It has been given both cognitivist-leaning realist

[4] Of course, one may also wonder whether this move would commit the naturalistic fallacy, be open to some sort of metaphysical open question arguments, or be in other ways problematic *qua* some form of naturalistic reductionism (Moore, 1903; cf. Enoch, 2011*b*, as cited in chapter 3, section 3 above). But there are also many answers to those questions, and this is not the place to discuss background assumptions.

interpretations (Copp, 2013; Smith, 1999; 2017) and expressivist-leaning anti-realist ones (Gibbard, 1999; Ridge, 2018). Moreover, more often than not, constitutivism is also taken to occupy minority metaethical positions such as relativism (Street, 2012; Velleman, 2013) or constructivism (Korsgaard, 2003; Street, 2008; 2010; Tubert, 2010). Simultaneously, Korsgaard holds that traditional metaethics is 'boring' and therefore should be avoided (cf. Korsgaard, 2003), whereas critics argue that she misunderstands metaethics, and that constitutivism therefore is inadequate (Hussain and Shah, 2006; 2013). We may call the latter charge *the metaethical problem*.

I believe constitutivism, as a general strategy, is compatible with several different metaethical interpretations; in fact, constitutivist moves may well count in favour of several theoretical frameworks. Hence, constitutivists need not decide what metaethical positions to take just *qua* constitutivists. However, different types of constitutivism lend themselves better or worse to different frameworks, and even do different kinds of jobs in them. I shall, therefore, respond to the metaethical problem by outlining how constitutivism works when coupled with some different metaethical views, along with suggestions for why it might be attractive given those frameworks. I start by discussing the main cognitivist and expressivist metaethical theories, and then conclude by saying something about the relation between constitutivism and more oddball -isms like relativism and constructivism.

Before I start, however, I want to note that my remarks here are somewhat unnecessary if one buys into the second positive motivation for constitutivism from above. It might be that constitutive norms just follow from truths about some aspect of agency. If that is right, there is little to do bar accepting them, potentially forcing one to revise more traditional metaethics. Hence, everyone should be either ok with, or worried about, constitutivist results.

But I shall also be more metaethically informative. To that end, I shall assume, by convention, that metaethical realism involves the claims that (*i*) moral language is cognitivist, and that (*ii*) there is a moral reality that this language refers to or describes. Here, constitutivism primarily does jobs in moral metaphysics and epistemology, as per the motivations for my view in chapter 1, section 3, and section 1 above.[5] Insofar as it has normative impact, that is usually secondary to its metaphysical and epistemic roles.

[5] It *may* also serve a normative role, as per the first-order motivations for constitutivism I mentioned in chapter 1, section 3. Whether it does so or not depends on other commitments.

More specifically, constitutivism shows that some properties are normative via some properties of features constitutive of some aspect of agency. There are different kinds of relations here on different views – usually explanatory – and constitutivism is compatible with views about these relations ranging from identity-reduction (as per reductive naturalism (cf. Schroeder, 2007)) to constitution (as per some forms of non-naturalism (cf. Shafer-Landau, 2003)). And then one can tell exactly the same epistemological story as the one I told in section 1 above.

Most versions of constitutivism do *not*, however, posit unexplainable (e.g. irreducible) normative properties. Hence, they are incompatible with views that do. Moreover, they do not explain normative properties by appealing to features of other things than some aspects of agency.[6] Views with these types of commitments may, at best, be hybrid forms of constitutivism, as constitutivist explanations only fundamentally appeal to properties of features constitutive of aspects of agency.

But, of course, one may not be a realist. One may deny either (*i*), i.e. that moral language is cognitive, or (*ii*), i.e. that there is a moral reality, or both. Denying (*i*) usually comes with an endorsement of expressivism about moral language, but whether there is a moral reality independently of our linguistic practices is strictly speaking a secondary question – there may or may not be one – so one need not necessarily deny (*ii*) at the same time as one denies (*i*). This matters because different routes here will give constitutivism different roles.

Assume, first, that one denies (*ii*) and becomes a moral error theorist. It might seem like constitutivism is incompatible with this view, for constitutivism is supposed to respond to the type of scepticism error theories involve. But this need not, strictly speaking, be the case. One may be a constitutivist about some normative phenomenon, e.g. normative reasons, and deny that there are any moral reasons. (This seems to be at least Joyce (2001)'s view, since his theory of reasons is Smith's, and that is a theory I have called constitutivist in chapter 6.) So constitutivism is, in one sense, compatible with error theories about moral norms.

This is an interesting result: it seems like constitutivism could serve to undermine the claim that there are moral reasons or facts (if these entail moral reasons) if conjoined with suitable premises. Nevertheless, one cannot be an error theorist and hold that moral constitutivism is successful. For if moral constitutivism works, there are moral facts or

[6] Schaab (2019), ch. 1, holds that what he calls 'Kantian constructivism' – which I would consider a species of constitutivism – is committed to *source internalism* about normativity, whereas source externalism is ruled out. This means, roughly, that the source of normativity has to come from the agent.

properties. Hence, depending on what one's form of constitutivism shows, one may end up either an error theorist or a realist.

Assume, instead, that one denies both (*i*) and (*ii*). For simplicity, I shall assume that anyone who denies both (*i*) and (*ii*) is an expressivist.[7] The paradigmatic contemporary expressivist is, probably, a quasi-realist (e.g. Blackburn, 1998; Gibbard, 2003).[8] Quasi-realists think that the function of moral language is to express attitudes, but that such expressions are truth-apt via a deflationist conception of truth, not in virtue of referring to or describing an independent reality.[9] They mostly treat talk of moral metaphysics – other than denying that there are moral properties in a realist sense – as talk that is internal to moral discourse, carrying no deeper implications.

On quasi-realist expressivist views, the role of constitutivism differs from the role it has on cognitivist views. This is just because of this lack of metaphysical import. There is no need to solve metaphysical queerness worries here, for quasi-realists do not believe in realist-style normative properties. But, simultaneously, constitutivism tends to generate a stable point in the middle of sets of moral judgements, as I argued when discussing the method of reflective equilibrium in section 1 above. Because the norms constitutivism explains turn out to be inescapable for agents (on most versions of the view), even though agents also can make other judgements according to quasi-realists, these other judgements will have to be made alongside the constitutivist ones. Constitutivism therefore seems to provide some moral norms a central role in agents' sets of moral judgements (cf. Ridge, 2018). This has both epistemic and first-order normative consequences. Constitutivism may help expressivism by anchoring normative reflection in virtue of providing both epistemic and first-order moral *stable points*.

Moreover, these moral stable points need not be understood as judgements about concrete moral questions. They can be; for example, if judging someone morally responsible requires us to take them to have a reason to act morally, then we can show how they have such a reason on constitutivist grounds. However, it might also be the case that one first-order moral stable point is that the source of the normativity of first-order judgements lies in agency. For example, perhaps we have reason to value our agency as a

[7] Doing so, I shall ignore quietist or fictionalist views that possibly should be interpreted as denying (*i*), (*ii*), or both. I am not sure about how they should be interpreted.

[8] Hence, I shall also ignore discussing old-fashioned forms of expressivism like emotivism.

[9] What is a deflationary conception of truth? There are many alternatives (cf. Lynch, 2001). But the most important point is that '*p* is true' serves as a kind of linguistic device for conveying assent to *p*; '*p*' is true iff *p*, where '*p*' is expressed in a meta-language and *p* in an object-language. Then we can speak of the truth of '*p*' without talking about correspondence, coherence, ideal consensus, or some other substantive truth-property.

part of a Kantian moral framework. Then that framework may receive a deeper, yet still possibly first-order, justification if the value is *grounded* in agency itself. Non-expressivist philosophers might treat the question of ground as (primarily) metaphysical, not as a first-order ethical question, but quasi-realists may instead be inclined to treat that grounding-question as a first-order ethical question. Along with enjoying the epistemic role constitutivism can serve, quasi-realist expressivists may therefore be more open to the normative motivation of constitutivism than I am (cf. chapter 1, section 3, esp. footnote 8; cf. footnote 5 above).

However, not all constitutivist views need be well-suited to quasi-realist normative sensibilities. If a quasi-realist is, for example, a reasons externalist, or wants to hold that there are stance-independent moral facts, she need not be attracted to the reasons internalism I have defended in this dissertation. To be fair, my kind of reasons internalism *may* be construed as a first-order normative view that she could endorse, but it is much more metaphysically formulated, and much less normatively assuming, than most forms of reasons externalism would be according to quasi-realists. Instead, the quasi-realist may instead be much more attracted to a more direct type of constitutivism, such as Korsgaard's, which has stronger first-order moral and rational implications (cf. Gibbard, 1999; Ridge, 2014; 2018).[10]

An interesting type of compatibility between expressivism and constitutivism appears, however, if the expressivist is a hybrid expressivist. Quite generally, expressivists need in fact not deny either (*i*) or (*ii*). Hybrid expressivists hold that moral language both serves to express judgements and report them – hence the hybridity. If that is right, a possibility opens up. The hybrid expressivist could think of constitutivism as doing metaphysical (as well as epistemic and first-order moral) work *because* the cognitive side of moral language is still in touch with moral properties. This means that hybrid expressivists can think of constitutivism like realists do.

Admittedly, Michael Ridge – a leading hybrid expressivist – argues that constitutivism provides very significant first-order benefits for quasi-realist expressivists for the reasons I mentioned when discussing quasi-realism (Ridge, 2018). However, he also criticizes standard forms of reasons internalist constitutivism (which he calls 'subjectivist'), holding that they oscillate between the *too many* and *too few reasons* versions of the *EXTENSION* objection I discussed in chapter 6, section 2. Constitutivists, he

[10] Admittedly, another option could be to treat my kind of ideal agent as a Gibbard-style hyperplanner (cf. Gibbard, 2003). One would have to do some translation work to make the frameworks coherent, however.

seems to think, would benefit from becoming quasi-realist style reasons externalists, just like quasi-realists might benefit from the stable points provided by constitutivists.

But I replied to the *EXTENSION* problem in chapter 6, section 2, so it is unclear why reasons externalism would be better than internalism. And as it is possible to combine hybrid expressivism with realist-style moral properties more straightforwardly than quasi-realistically, Ridge does not need to go quasi-realist as opposed to more fully realist about the cognitivist side of his hybrid view. He could stick with a version of hybrid expressivism, but also accept some decent form of reasons internalism (e.g. *DESIRES INTERNALISM*), and then get a moral reality that the cognitive aspect of moral judgements may refer to or describe. If my arguments are at all plausible, this view should present an attractive option even for a hybrid expressivist.

So much for realism, error theories, and expressivism. Constitutivism seems compatible with them all, in one way or another. This leaves us with more oddball metaethical -isms, such as relativism and constructivism. Strictly speaking, both of them are orthogonal to the main theoretical contenders in the field; there can be relativist cognitivism and expressivism, as well as constructivist cognitivism and expressivism. But because constitutivism sometimes is run together with these -isms, it is worth saying something about how it relates to them.

First, by 'relativism', I here mean truth-relativism, whether speaker- or assessor-based, plus the extra assumption that not all relevant agents will be assessed similarly, so different moral truths hold for different agents.[11] Relativists of these sorts may be either cognitivist or non-cognitivist and have no particular difficulty with constitutivism – though one may, of course, want to shy away from relativism for other reasons. It is not obvious how relativist views can preserve the universal prescriptivity of morality, for example. Nevertheless, some forms of constitutivism might even be thought to generate relativist conclusions. Velleman (2013) thinks this about his own type of constitutivism, and Street (2008; 2012) holds a view which may do so, too.

Finally, 'constructivism' can be given two main interpretations. The first is ontological, based on the assumption that a moral fact (or property) is constructed through some constructive procedure. I personally prefer treating this kind of constructivism as a form of naturalist realism, but regardless of how one construes it,

[11] Hence, I do not count the 'relativism of modal distance' above as a relevant form of relativism. It is not committed to truth-relativism, and even if it had been, 'all relevant agents' would be all agents who can perform means-ends actions, and they are assessed similarly.

constitutivism functions exactly like it does for realists on such views, so there is not much more to add here.

But one may also be a constructivist about truth, which is a view I tend to equate with holding that truth is constructed by some procedure, e.g. ideal inquiry. Such views can come in either cognitivist or expressivist shape, or in versions that seem to combine elements of both (e.g. Korsgaard, 2003). Here, constitutivism may still come to serve all the three main motivations for it; in fact, many think of constitutivism as a form of constructivism along these lines (Korsgaard, 2003; Tubert, 2010; Schaab, 2019, ch. 1). Certainly, there is no inconsistency here.

Hence, to summarize: constitutivism seems compatible with most metaethical theories. It is, surprisingly, even compatible with some weak forms of non-naturalism. But it is *not* compatible with views according to which normativity is primitive or otherwise external to the features of agency. Such views may at best be hybrid constitutivist ones, taking some norms to be explained by agency and some by other means. But not being compatible with such views is hardly a problem – constitutivism has often been raised as an alternative to their alleged failures.

(5) Conclusion

I have defended a version of *FORMAL CONSTITUTIVISM$_{PM}$*. It seems generally justifiable in virtue of its ability to handle sceptical worries (section 1) and explanatory strengths (section 2). It does better than other constitutivist views about moral norms when it comes to the problems of substantive assumptions and adequacy as well as the shmagency objection (section 3), and it is compatible with the solutions to other problems that I have presented (section 4). I conclude that, even though I have not conclusively defended my form of constitutivism, it is a serious metaethical option which can benefit many metaethical frameworks.

Appendix A: Objections to *HTM*

In this appendix, I extend my argument for *HTM* (in the form of *CENTRAL HUMEANISM*). I will do so by replying to objections to *HTM*. My aim will, however, be limited. I do not have the space to develop responses to all objections to *HTM* in depth, nor even to respond to all objections.[1] Instead, I shall suggest some answers to some key objections to the theory.

In section 1, I discuss how the components of *HTM* – in particular, desires – work. In section 2, I discuss which explanations *HTM* can provide. And in section 3, I discuss the relation between *HTM* and acting for reasons. I conclude in section 4.

(1) *HTM* and Its Components

I have defended:

> (*CENTRAL HUMEANISM*) A necessary feature of what makes one central class of events intentional actions is that a belief/desire-pair, suitably linked up, non-deviantly is part of the cause of the events in question.

Furthermore, in chapter 5, I argued that the belief/desire-pairs that cause and make events paradigmatic actions must be linked up by *IR_{HUMEAN}*. This means that the key constituent features of *HTM* are beliefs, desires, and *IR_{HUMEAN}*. But we may ask how we should understand these features – in particular, how desires work is very controversial.

Now, I have taken desires to be functionalist-style mental states with, at least, the function of causing action, in particular when linked up with beliefs via the instrumental principle. In fact, for Humeans, action-motivation is a *necessary* feature of desires, and they themselves necessarily feature in the etiology and constitution of paradigmatic actions.[2] The latter two properties are the only ones I need to get my argument going.

However, it is simultaneously plausible that desires can have all kinds of additional features. For example, they can be occurrent or not, and it is plausible that they are at

[1] In particular, this means that I will not spend a lot of time on potential counterexamples to the extensional adequacy of *HTM*. But some of these are discussed elsewhere (e.g. chapter 5, section 7), and many other sets of counterexamples – cf. references in chapter 4, section 1, footnote 6 – are fully compatible with *CENTRAL HUMEANISM*, for it does not rule out that there can be actions of non-fundamental kinds.

[2] This leaves it open whether desires might motivate without beliefs or whether there might be other mental states that motivate. I take no stance on these questions here.

times knowable, and at times not. And I have argued that desires perhaps play some sort of epistemic role. Knowledge of our desires can let us know our reasons, as a kind of heuristic.[3] And I have even suggested that it is possible that desires show their objects in an attractive light.

Why these properties? As usual, my methodology involves listing key features of some phenomenon, and then trying to see how they can be explained (cf. chapter 1, introduction). This is what I have done here too. Desires seem to at least motivate action, come in different degrees of strength and knowability in experience, teach us something about our reasons (at times), and sometimes make their objects seem attractive. So desires seem apt for having at least the properties I have mentioned.

But one may wonder whether this is an accurate characterization of desires. I have not explicitly considered competing views, and there are many of them. For example, some take desires to be primarily epistemic in directing us towards the good (Oddie, 2005; Stampe, 1987), and some take them to be judgements of it (Scanlon, 1998, ch. 1), often in ways where the desires end up reduced to beliefs (cf. Schroeder, 2015). So why should we think that desires motivate?

The answer is that motivation is a deep feature of desires in ordinary folk psychology, as well as in the social sciences (cf. chapter 4, section 2). Any view that does without a desiderative conception of motivation loses out on the explanatory role they seem to possess. So we should maintain a motivational conception of desires for the same reason that we should stick with explanations involving beliefs and desires – they provide good explanations of why we act (and, I have argued, of what makes events actions).

Moreover, the epistemic role that many indicate that desires can play can still be captured on the motivational view. It is possible to use them to gain normative knowledge as a kind of heuristic, even though that is not one of their functional roles. This is because reasons are grounded in idealized desires (cf. chapter 6), so there is an intimate connection between reasons and desires. This move gives us the right result here: knowledge of reasons through desires is often murky, for desires are famously often and easily misleading, and this is nicely captured by the Humean view.

[3] More controversially, perhaps '[desires cause] pleasure and displeasure when we have changing beliefs or vivid representations concerning [their] satisfaction. [They direct] attention towards things we associate with its object. [These] effects are amplified when we have vivid representations that we associate with [their objects, and] intrinsic desires don't change through reasoning' (Sinhababu, 2017, p. 22). If desires have these properties, they can probably explain many potential counterexamples to *HTM*. But I shall not assume that they do.

Another challenge to the view that desires fundamentally are motivational is that they might not seem like good contenders for motivating action because they are based on something else from the start themselves (Hornsby, 2004; Nagel, 1970, pt. II). Or, alternatively, perhaps they cannot explain actions if they are not motivated by something else (Dancy, 2000, ch. 4). In particular, many think desires have their bases in reasons for action. We act for reasons for action based on (external) facts rather than desires, and explaining actions in terms of desires is at best secondary to that.[4] Sometimes, this point is put by saying that desires are *motivated* by reasons.

This point seems extensionally inadequate. Often, one brutely desires some things, and deeper rationalizations seem overintellectualizing in these cases. Why do I want to eat spaghetti Bolognese (but not mac and cheese)? I do not find spaghetti Bolognese tastier, I am familiar with the taste of both, I find them equally cool or uncool, etc. So what else than a desire could explain my preference? And as they have all kinds of properties – as I just argued – why would that not be enough to explain potential actions?

Nevertheless, I concede that our desires often appear motivated – because even though they need not be, they often are. Humeans can allow for that by appealing to how desires can be motivated by other desires (or desire-based reasons). If I desire to eat tasty food, and that grounds a reason to eat tasty food, and I find spaghetti Bolognese tasty, then plausibly I have a(n instrumental) reason to eat spaghetti Bolognese – and *ipso facto* a kind of support for my desire to eat spaghetti Bolognese. Such cases can explain the intuition that desires sometimes appear to be motivated.

(2) *HTM* and Its Explanations

It is often argued that the kind of causal explanations *HTM* can give of our actions is insufficient to capture how we cause actions. I shall discuss two types of such problems.

First, there is the problem of deviant causal chains. This is a famously vexed problem for causal theories of action – and, not least, for Humeans (Davidson, 1980*b*). I follow Mayr (2011, ch. 5) in taking there to be at least two versions of the problem: the *problem of antecedent deviant causal chains*, and the *problem of consequential deviance*. In the former case, an agent causes an event via her belief/desire-pair (or intention), but does so in the wrong way. Perhaps she intends to shoot someone and holds her finger on a gun trigger,

[4] Yet another version of this argument suggests that desires or actions aim at the good, and hence cannot explain reasons – it is the good that does (cf. Frey, 2019; Railton, 1997). But desires can aim at everything (Stocker, 1979). Moreover, the responses in the main text apply to that version of the argument too.

but this unnerves her so much that she starts shaking, which leads her to shoot that very person. In the latter case, perhaps she shoots and misses, but her shot ricochets of a nearby rock, thereby hitting the person. In both cases, the agent causes events in the wrong way, so it seems problematic to count them as actions.

Versions of this problem also trouble analyses of various other phenomena, e.g. knowledge (cf. Enç, 2003). At least some Gettier cases have the same structure as the problem of deviant causal chains. In those case, one has a justified true belief which is only luckily true in the same way as one only luckily achieves one's aim in 'actions' based on deviant causal chains. For example, I may believe that Victoria the skilled archer will hit the bull's eye because she usually does, she shoots, but the wind catches the arrow, blowing it in one direction, but then back towards the target – and it hits the bull's eye. In their general form, I have previously called cases like these instances of 'Reparatory Luck'. They are such that the world unsettles one's aim, but then also resettles it by accident (Leffler, 2016).

The fact that cases of reparatory luck generalize is interesting. The problem of deviant causal chains has long been raised against all versions of the causal theory of action, just as it has against various theories of knowledge (with Gettier cases). Yet few have wanted to deny that we can have knowledge at all because of Gettier cases, or even that we cannot explain knowledge. Rather, philosophers have revised their analyses or explanations of knowledge. This indicates that we can take inspiration from epistemology in our solution to the problem of deviant causal chains in philosophy of action.

There are several suggestions for how to solve the problem in the literature. One which is common in both epistemology and philosophy of action, and which would be natural to take up for me, is a sensitivity approach. Here, an agent's ability to act depends on responding counterfactually sensitively to shifts in events over possible worlds. The idea in the action case is that one only forms and enacts instrumental desires that are sensitive in such a way. This is how Smith (2012*a*) solves the problem of deviant causal chains, and when I defend the role of instrumental rationality in action (cf. chapter 5, section 3), I appeal to the same type of capacity for rationality as he does in that paper. So I could extend it to solve the problem.

But while going down that route is an option, it would still be a contentious endeavour (cf. e.g. Mayr, 2011, ch. 5). Instead, though I cannot develop it in detail here, I would like to introduce an alternative approach which draws on the action disjunctivism I use to solve the problem of bad action in chapter 5. For one could argue that belief/desire-pairs causing events through deviant causal chains are cases of bad action. They fail to cause events with sufficiently much control to count as successful actions, but still cause events, and hence may belong to a different kind of actions than successful cases (cf. Lord, 2018, pt. III).

This move is parallel to one made in some forms of epistemological disjunctivism. For Williamson, knowledge is the most general factive mental state, and to believe that *p* is to treat *p* as if one knows it, so non-factive belief is failed knowledge (Williamson, 2001, ch. 1; cf. chapter 5, section 7). On this view, Williamson can treat true beliefs suffering from Gettier causal pathways as beliefs that fail to be knowledge. Analogously, we may treat deviantly caused cases of 'actions' as actions of a kind that fail to be successful ones. Events caused by ordinarily action-causing mental states via deviant causal chains are, in my terminology, not paradigmatic actions, but rather events we may treat as if they were paradigmatic actions even though they are not, for they have deviant mental causes. And while paradigmatic actions are the fundamental kind of action, such events can be treated as belonging to a non-fundamental kind of action.

It might be replied that failed actions, via deviant causal chains, are not actions at all. However, looking at the examples I used to introduce the problem, it seems like there is something very much *like* action going on here. In both the cases of antecedent and consequential deviance, it does not seem wrong to hold that the agent, in some sense, shoots her target. In both cases, it is the agent who initiates movements based on her belief/desire-pairs, though these states end up causing the event in the wrong way. So it is not very far-fetched to treat these cases as actions of a non-fundamental kind – they still have all the components of action, minus the right causal pathway. Hence, disjunctivism shows promise when it comes to explaining cases of deviant causal chains. It is worth investigating further whether it might be able to explain all possible cases.

The second major problem having to do with causation is the problem of action without movement. More often than not, *HTM* is defined in terms of beliefs and desires causing some bodily movement. Yet many have questioned whether it can explain actions where the agent does not seem to initiate any form of movement. There are, at least, three broad classes of potential actions of this kind. They are:

> (*DELIBERATE RESTRAINT*) Where the agent decides actively *not* to move.
>
> (*OMISSION*) Where the agent decides not to initiate some course of action.

> (*MENTAL ACTION*) Where the agent acts only mentally, hence not causing bodily movements.

However, I did not define *CENTRAL HUMEANISM* in terms of initiating bodily movements. It is formulated in terms of belief/desire-pairs causing *events*, so one need not cause movements here.[5] Yet a puzzle remains. How does the agent cause *these* events? After all, should one not have some form of positive causal impact, altering the world, when one causes something? Perhaps one has that in cases of *MENTAL ACTION* – one causes mental events – but one need not in cases of restraint or omissions.

The answer to this question can be found if we disentangle two forms of causation. There is *change-initiating causation*, where *F* causes *G* by initiating some new chain of events, featuring *G*, and *sustaining causation*, where *F* causes *G* because *F* sustains the existence of *G*, which already exists (cf. Lord, 2018, pp. 135-143).[6] With this distinction in mind, we can treat *MENTAL CAUSATION* as a case of change-initiating causation. But we can also make sense of *DELIBERATE RESTRAINT* and *OMISSION*. In these cases, the agent's belief/desire-pair *sustain-causes* already occurring events because she has decided not to alter them, which is part of the cause of why they remain the same. So, using this distinction between kinds of causation, *HTM* is able to explain actions where one does not change-initiatingly cause events, but rather sustain-causes them.

(3) *HTM* and its Reasons

The third and final set of explanatory problems concerns the relation between causal and rationalizing explanations. I shall focus on some major ones I have not already answered.

First, one may wonder how desires are part of rationalizing explanations of actions. A traditional view for Humeans is to hold that belief/desire-pairs rationalize an action by constituting a motivating reason for it.[7,8] This motivating reason rationalizes the

[5] I want to thank Björn Petersson (in conversation) for the events suggestion.

[6] This type of sustaining causation differs from the type of sustaining causation that may explain the sustained functioning of some system over time by responding to negative feedback loops (cf. Mayr, 2011, ch. 5). This one need not sustain an entity over time or respond to feedback loops at all.

[7] I find it natural to talk in terms of constitution, but perhaps the reader prefers some other relation here. Nothing turns on this.

[8] Even this may be a slight simplification. With *CENTRAL HUMEANISM*, let alone action disjunctivism, I have allowed for the possibility that there can be different kinds of action. Are non-Humean actions performed for motivating reasons? Suppose they are not. Then there is no worry for *HTM*; all actions featuring motivating reasons are still caused by belief/desire-pairs. But suppose they are. Then maybe the class of motivating reasons would have to be expanded to feature other motivational states than

action by making it intelligible, where 'intelligibility' means something like how the action makes sense to the agent, e.g. because it is an appropriate outcome given what the agent takes into account before acting (O'Brien, 2018; cf. chapter 4, section 1). On the other hand, a motivating reason may or may not be congruent with our *normative* reasons for action. This means that the motivating reason is not a reason in a normatively very significant sense of the word, so to avoid confusion, I usually prefer to talk about belief/desire-pairs rather than motivating reasons.

It is very plausible that a belief/desire-pair can make an action intelligible, however. A desire both sets an end for the agent, showing what she may take into account, and sometimes works as a heuristic for our knowledge of our normative reasons, hence indicating, to the agent, that she has reason to do what she desires. Accordingly, acting on a desire usually makes sense to the agent. Of course, this does not necessarily entail that any particular motivating reason must be in line with a normative reason. Not even paradigmatic actions need be motivated by normative reasons. But that has little to do with whether or not they are motivated by motivating reasons.

However, Jonathan Dancy (2000) has objected that motivating reasons and normative reasons both should be counted as considerations that count in favour of action. The best account of reasons explanations, he thinks, is one according to which there is no substantive, metaphysical, difference between explanations in terms of motivating and normative reasons. Humeans like me believe that motivating reasons are psychological, whereas normative reasons are facts grounded in desires (cf. chapter 6). But Dancy holds that reasons are external to the agent; they are relations holding between her and states of affairs, and there is no substantive difference between explaining action in terms of such reasons or her mental states. It follows that there is no deep distinction between motivating and normative reasons; they are the same external reasons, thought of in different ways.

However, *because* motivating and normative reasons should be metaphysically distinct, Dancy's framework fails. Why? We should want to say that there can be cases where motivating reasons and normative reasons need not go together. This is because motivating reasons are the considerations in the light of which actions are intelligible to the agent performing them, whereas normative reasons have normative force. And this, in turn, is why motivating reasons can feature desires with bad contents, whereas

belief/desire-pairs. I suspect there are intuitions pushing in both directions, so at present, we need not decide.

normative reasons cannot (cf. Smith, 2003*b*, for similar considerations). But because it is possible to act intelligibly for no normative reason, or even for normatively bad reasons – which are not normative reasons – it is plausible that we should think of motivating reasons as different from normative reasons. Hence, the distinction between normative and motivating reasons should be upheld.

(4) Conclusion

I have now suggested some ways to defend *HTM*. In section 1, I claimed that a fairly substantive conception of desires explains much about them, and also shows how they can play the right role when motivating actions. In section 2, I claimed that *HTM* can handle objections about causation because deviant causation either is something it is sensitive to or causes non-fundamental actions, and because actions without movement can be explained by appeal to a distinction between two kinds of causation. And in section 3, I claimed that belief/desire-pairs plausibly have very different roles from normative reasons, and this difference should be sustained, not least because we can act for normatively neutral or even bad 'reasons'. No doubt this defence remains provisional and needs expansion, but that means that it is fruitful. Humeanism ain't dead yet.

Appendix B: Beyond Instrumentalist Constitutivism about Rationality?

In chapter 5, I argue that a principle of instrumental rationality is constitutive of paradigmatic agency. I even claim that this principle is normative. But even though I defend it from some pertinent criticisms, it remains extremely controversial. In this appendix, I try to defend my principle against alternatives in the burgeoning literature on the topic. The defence will still have to be fairly cursory, however, but I will at least show that it is a decent contender in comparison with other theories in the literature.

I start in section 1 by discussing how it does better than other types of constitutivism about practical rationality. In section 2, I discuss scepticism about rationality. In section 3, I discuss alternative theories, according to which practical rationality should be understood either in terms of theoretical rationality, agents' capacities to respond to reasons, or agents' virtues. I conclude in section 4 with a methodological *caveat*.

(1) Other Forms of Constitutivism

There are many types of constitutivism in the literature about rationality, even beyond mine. First, there are Kantian views. Here, the idea is to treat the categorical and hypothetical imperatives as constitutive principles of agency (i.e. as *C* of *A** in *FORMAL CONSTITUTIVISM*), and *also* as norms of rationality.

For discussion of *CI*, I refer the reader to what I said about Korsgaard's view in chapter 2, section 2. But what about *HI*? The core argument here is that the relevant aspect of agency (i.e. *A**) is willing, and the hypothetical imperative is constitutive of having a will. To will φ, one must follow the hypothetical imperative via taking some means to φ (on pain of irrationality), and it is for that reason that willing differs from merely desiring (cf. Korsgaard, 2009, pp. 68-70). But I do not see what willing would have to be beyond having an instrumental desire. Instrumental desires seem able to play the same role as 'the will' is supposed to do. So the Kantian views fall back into the Humean one I defended in chapter 5.

But there are also other Humean views. On one kind of Humean view, taking means to ends is instrumentally rational because it is constitutive of some mental states – such as desires, intentions, or some subset of these that count as 'ends' – to take means to ends (cf. Fink, 2014; Goldman, 2011, ch. 2; Smith 2012*c*; Railton, 1997).[1] On the first of these views, the relevant aspect of agency *A** is desires, the constitutive feature *C* is the instrumental principle, and the normative phenomenon *P* – instrumental rationality – might be explained by how it is constitutive of desires when these are fully functional without reference to a separate principle or capacity of *IR*.

But the view that the instrumental principle is 'built into' desires does not seem compatible with the response to the problem of the disappearing agent I developed in chapter 5. If one were to hold that it is, one would have to say that the agent stands behind an action in virtue of desiring it, but the agent need not stand behind the action as something over and above it if the action is just caused by her desires (and beliefs), or even actively cause it, because there need not be anything active about having any particular desires cause events. Hence, a Humean view that takes the instrumental principle to be constitutive of desires (or other ends) is worse off than mine. It lacks a compelling alternative solution to the problem of the disappearing agent.

What if the mental state which is partially constituted by rationality is something else than a desire, such as an intention? Some philosophers – Humean or other – would like to explain requirements of rationality in terms of intentions (e.g. Kauppinen, forthcoming; Mylonaki, 2018, and references therein). I do not, fundamentally, have a view about whether that might work; I am silent on the nature of intention. But insofar as an alternative explanation of instrumental rationality can be given without appealing to intentions, as per *PARADIGMATIC AGENCY*$_{IR}$, these explanations seem redundant. This leaves such views *prima facie* unnecessarily unparsimonious.

There are also views that cannot be neatly placed in the Kantian or Humean traditions. These views tend to be holistic, in the sense that they attempt to explain several rational principles. The first of these we can call 'proper functioning' views. Such views are defended by Michael Bratman, Michael Smith, and others (cf. chapter 3, section 5). According to proper functioning views:

> [R]ational requirements describe the proper functioning of certain distinctive systems that constitute the special kind of

[1] Goldman's view about instrumental rationality focuses on self-defeating intentions, but his broader conception of irrationality as self-defeat is more complex. The core idea is that self-defeat – both on the practical and theoretical side – grounds various demands of rationality. The self-defeatingness is sometimes a feature of distinct mental states, but at times also of actions.

This means that his view fails spectacularly as far as shmagency is concerned, however. It is very easy to have other mental states than beliefs, desires, etc., and not be self-defeating, and therefore there is nothing obviously normative about Goldman's rational standards.

> agency that we possess. (…) The basic idea is as follows. Our psychological attitudes are to be understood as elements in systems that serve certain functions. Rational requirements describe what is necessary for our attitudes to serve these functions properly. (Southwood, 2008, p. 22)

There are many different proper functioning views, but they all feature rational requirements constitutive of psychological *systems*, not just of individual states inside those systems. According to Bratman, there can be different kinds of agency, but a kind of autonomous, cross-temporal, agency, constituted in part by certain principles of rationality, has special value for us. And according to Smith, agency is by itself a goodness-fixing kind, with its own functional standards. But regardless of which view we go with here, proper functioning views count as constitutivist. The norms of rationality *P* are normative in virtue of properties of the constitutive features *C*, i.e. the proper functioning of an aspect of agency *A**.

As Southwood notes, however, this form of constitutivism seems vulnerable to shmagency objections. It is unclear why we would have to be some complex kind of agents; it is quite possible to be an agent that does not function fully. But if one instead, like Bratman or Smith, would like to defend the normativity of rationality by appealing to the normative inescapability of the principles of rationality, one would have to engage in a significant amount of normative theorizing to show why the principles of rationality would be normatively inescapable (cf. chapter 3, section 6).

Might one construe my view of instrumental rationality, too, as a proper functioning account? I think so. I have argued that instrumental rationality serves a role in making our psychological attitudes cause actions (cf. chapter 5, section 3). But is that an objection to my view? No, for the system view I defend is more minimal than Bratman's or Smith's – I only defend a system needed to initiate individual actions, not a system featuring temporally extended and autonomous agency or agency construed as a goodness-fixing kind. And my kind of agency is plight inescapable, not normatively inescapable.

Shmagents do, to be fair, appear relative to all our views. But unlike their shmagents, mine are harmless. It is plausible to think that paradigmatic human agents can act via IR_{HUMEAN} since that is a minimal systemic commitment, whereas the kinds of norms they want to defend seem more akin to norms of agency *par excellence*, not of paradigmatic actual agents. If their norms are inescapable at all, they are normatively inescapable, but then they run into the underdetermination version of the shmagency objection, as per the above. But I do not (cf. chapter 7, section 4; chapter 8, section 3).

Perhaps the most substantive form of constitutivism in the contemporary literature about rationality is, however, Southwood's own (Southwood, 2008; 2018*b*; cf. Broome, 2008; Coons and Faraci, 2010; Levy, 2018 for criticism). He has argued that requirements of rationality are normative in virtue of being constitutive (*C*) of having a first-personal standpoint (*A**). Here, a standpoint is 'constructed out of our particular beliefs, desires, hopes, fears, goals, values, and so on, and relative to which things can go well or badly. Our standpoints describe what matters to us; they are ones in which we are invested' (Southwood, 2008, p. 26). And because the requirements of rationality are constitutive of our standpoints, they are normative. This grounds a special kind of first-personal normative force for them.

Unfortunately, Southwood's view is susceptible to the main objections that I have been pressing against constitutivism throughout this dissertation. First, it is directly implausible in its assumptions about agency, generating the problems of adequacy and substantive assumptions. No reason has been presented to think that there is a distinct kind of first-personal normative force.

Second, there are shmagency-style worries. It does not seem like we, descriptively, necessarily have his first-personal standpoints. Perhaps we can be easily swayed by fashion, e.g. the winds of political rhetoric, attaining or retaining new desires, values and dreams, or perhaps we are just inclined to change our minds easily. If so, there is no interesting sense in which we need to have standpoints with things that matter to us or in which we are invested. We could just have shmandpoints.

Southwood has developed his view, however. In Southwood (2018*b*), he has suggested that the norms of practical reason apply and have force in another way: they govern answers to the question of 'what to do', where that is not a question of what is required by deliberative agency, but rather truths that determine what 'the thing to do' is. The question of 'what to do', he adds, is the question one attempts to answer when one uses one's faculty of practical reason. It is not answered just by appealing to what is required by being an agent.

In the light of the last point, Southwood's new view does not seem constitutivist. But it can be reformulated in constitutivist terms. Perhaps the faculty of practical reason is (partially) constitutive of agency, and hence answering the 'what to do'-question is partially constitutive of agency. Alternatively, perhaps practical reason is an aspect of agency, whether or not it is constitutive of agency *tout court*. That is enough to generate a constitutivist version of his view.

Alas, the reformulated view has no response to shmagency-style objections either. Why should we not aim for 'the thing to shmo', and hence use some ability of practical shmeasoning instead of an ability of practically reasoning? (Of course, shmoing may be guided by something else – e.g. principles of shmationality – so it is not the case that it is wholly unguided.) At the very least there can be Uranians who have a faculty of practical shmeason that allows them to do that, making them shmagents as sophisticated as the Martians and Saturnians of chapter 3. This generates a version of the first shmagency worry – Southwood does not explain these norms for shmagents who plausibly should be subject to them. Nor is it clear to what extent the thing to do is one that fully utilizes, or utilizes all the norms of, practical reasoning. That gives rise to a new version of the second shmagency problem, too.

(2) Scepticism about Rationality?

Other types of constitutivism do not seem as plausible as the Humean version I have defended. But some philosophers have gone further still and denied the normative role of principles of rationality. This happens on two types of theories. Some philosophers, e.g. Kolodny (2005; 2007*a*; 2007*b*; 2008*a*; 2008*b*; 2009) and Raz (2005*a*; 2005*b*), argue that there are no normatively distinct requirements of rationality. Structural requirements of rationality indicate that we have failed to take our reasons into account, but these requirements are not by themselves normative. When we fail to follow them, that only means that we fail to take our reasons into account.

Other philosophers do think that rationality can be normative, but that it consists entirely of responding to reasons (e.g. Henning, 2018; Kiesewetter, 2017). These philosophers do not believe there are any principles of rationality; rather, practical reasons do all the normative and explanatory work, so they, too, are sceptics about normative principles of rationality, but about the principles rather than the normativity of rationality. But because of how similar this view is to the view I just mentioned in the last paragraph, I shall run the two types of scepticism together for present purposes.

Together, the sceptics I have mentioned have presented a battery of arguments against the idea that there are normative principles of rationality. Henning argues that we should understand conditionals of rationality (e.g. 'if you want to φ, you ought to ψ') as systematically ambiguous; they can either express psychological states on part of the agent who has them or content about what to do, and this accounts leaves us without any need to appeal to structural principles. Kiesewetter and Kolodny, moreover, present many different arguments over the course of many papers and books. And Raz focuses on instrumental rationality in particular. Nevertheless, their core point is that rationality does not provide any particular *reasons*.

It is not necessary to rehearse all their arguments here. Since my argument is constitutivist, I shall focus on their arguments insofar as they have bearing on constitutivism. First, Henning's linguistic argument has little to do with how we should think about the principle of instrumental rationality from my perspective. I have defended the principle of instrumental rationality as a psychological feature of paradigmatic actions and agency in virtue of its theoretical role in explaining how the agent stands behind actions. And even though I talked about some commitments about rationality from ordinary discourse when I presented the case for taking instrumental rationality seriously, the main argument in its favour is quite independent of our interpretation of conditionals in ordinary language.

When it comes to the more substantive arguments, Raz agrees with (some) constitutivists that irrationality may consist of some sort of psychological malfunctioning, yielding incoherent beliefs. But that is, he thinks, not a reason to be rational. Incoherence is not reason-giving; its role is, instead, to indicate that we have failed to take all our underlying reasons into account.

Kolodny is sympathetic to this account (Kolodny, 2008*b*), and he adduces several distinct considerations against constitutivism. First, he argues against the view that mental states would have intrinsic norms (independently of other mental states) (2008*a*; 2008*b*). I made the same claim in section 1 above, so here we agree.

However, he has stronger arguments still. In the same paper series, he also argues that principles of rationality are not constitutive of anything interesting because they do not lose their force even though one can fail to act on them in particular cases. Moreover, allowing that they may apply to subjects holistically, he argues that there can be variations in the extent to which one may lose one's agency by not acting on them, and that this should affect what one rationally ought to do. But, he adds, principles do not predict to what extent one may lose one's agency by failing to act on them, so the principles do not have much to do with the extent to which one maintains or loses one's agency. Hence, he concludes that the oughts of rationality do not vary with how far one is from losing one's agency. Finally, he claims that constitutivists may be able to show why we are committed to rational principles, but that that does not show that they give us reasons to follow them (Kolodny, 2005; 2007*a*). Rather, constitutivism can at best provide an explanation suggesting *that* we are obliged to follow them, rather than an explanation of *why* we are.

However, Raz and Kolodny have not taken the argument from chapter 5 into account. Raz claims that without explaining why ends should have impact on our thought and intentions, constitutivists leave it unclear 'how having ends is relevant to the well-functioning of our deliberative processes' (Raz, 2005*b*, p. 15). But the argument from chapter 5 explains why ends have bearing on actions. Ends, construed as desires, are part of instrumental reasons, which explain action.

Does Kolodny have more to say here? His first important point – about the possibility of error – is easily accounted for. That is what section 7 of chapter 5 is supposed to do. His second argument, about the variability of rational ought, is harder to assess. I have wanted to restrict myself to claims about requirements, leaving the semantics of 'ought' on the side. Nevertheless, I do not see why instrumental rationality would have to vary with actions. It applies to every action, not to agency holistically construed. Paradigmatic agents may have the capacity to be instrumentally rational without exercising it in every action.

Finally, I agree with Kolodny that rational requirements do not provide reasons for action. But the normativity of the instrumental principle is not supposed to be explained by its reason-providingness. It is a directive standard of success (cf. chapter 5, section 5; 6). Hence, my theory about the normativity of instrumental rationality is distinct from the views that Kolodny criticizes.

Here, one might think that I have begged the question. I *ought* to show how rationality is reason-providing, not how it is a directive standard of success. But it is unclear why I should have to concede that much to the sceptics. They do not spend much time trying to show that reason-providingness is the gold standard of normativity, whereas I have tried to show that the kind of normativity I have defended for instrumental rationality is one that is apt for it (cf. chapter 5, section 1; 5). Hence, I shall stick with my explanation.

In fact, in the light of my take on the features of the normativity of rationality, one may wonder why one should take reasons to be more normatively central than rationality. This point takes us to Kiesewetter's intricate (2017) view. While Kiesewetter barely discusses constitutivism, his picture differs substantially from mine; he thinks rational conditionals are normative because we have reason to follow them, but they have nothing to do with structural principles, because there are no such principles. He employs a battery of arguments to reach that conclusion.

As mentioned, however, I do not think the normativity of rationality has much to do with reasons. That part of his argument seems irrelevant in relation to mine. And this point about the construal of normativity also takes us to my point about centrality. Normativity aside, Kiesewetter's core case against structural principles of rationality in general rests on providing several arguments against both narrow- and wide-scope principles of rationality (cf. Kiesewetter, 2017, ch. 6). He prefers, instead, to explain the how rational principles appear to – but do not – exist by appealing to our reasons (Kieswetter, 2017, ch. 10-11).

I lack both the space and the theoretical inclination to engage with his arguments substantially here; I want to remain neutral on the wide-scope vs. narrow-scope distinction (cf. chapter 5, section 2). But I do want to emphasize a point about orders of explanation. Insofar as Kiesewetter's theory relies on explaining rationality in terms of reasons, we need a theory about how reasons work. I have tried to provide one in terms of idealized desires. Is my view more (or at least just as) plausible as his view, where reasons are unexplained?

It seems extremely hard to formulate a theory where reasons are prior to rationality without suffering from standard Mackie-style metaphysical and epistemological arguments from queerness (cf. chapter 1, section 3). Of course, one may try to reply to those arguments, but doing so is far from easy. Hence, my explanation of reasons in terms of idealized desires comes off as at least as attractive as Kiesewetter's more general framework, whether or not an acceptable theory of wide- or narrow-scope instrumental rationality has been developed at present.

(3) Other Approaches

Maybe, then, my type of constitutivism about rationality is better than other types, and maybe sceptical theories can be avoided. But I have not discussed other central positive theories about rationality in the literature. The first of these is cognitivism, according to which principles of practical rationality fundamentally have a theoretical explanation; they are standards of rationality for belief (Setiya, 2007; Wallace, 2001). The second view takes reasons to be prior to rationality, and then interprets the normativity of principles of rationality in terms of reasons-responsiveness (e.g. Lord, 2018; Parfit, 2011*a*). And the third view takes rationality to primarily be an aretaic or virtue-based concept, and then explains principles of rationality in such terms (Svavarsdóttir, 2008; Wedgwood, 2017, ch. 6). I shall discuss these views in this order.

First, cognitivism. The idea here is that practical rationality, including the instrumental principle, in one way or another can be explained as requirements of consistency on beliefs. Intentions (or, possibly, desires), cognitivists think, involve or include beliefs, so not taking what one believes to be the (best) means to what one intends involves having inconsistent beliefs.

My response to the intention view is the same as the one in section 1 aimed at others who appeal to intentions to explain rationality. I have no settled view about intentions, but if an explanation of instrumental rationality can be given without appealing to intentions, intention-based explanations turn out to be unnecessarily unparsimonious. And my response to the belief-as-desire view is that I defend the explanatory role of desires in appendix A above.

Second, there are reasons-responsiveness views. Writers who hold such views tend to want to explain requirements of rationality in terms of reasons (cf. e.g. Lord, 2018; Parfit, 2011*a*).[2] I am not sympathetic to this general strategy in the light of the worry I voiced against Kiesewetter's view in the last section. It is highly unclear to me how reasons should be understood unless one goes with a rationality-first view (like mine) which explains them from the bottom up.

Another reversal of explanatory roles is made in the last major contending view to my rationality-first one. That is the view where rationality is understood as a virtue. Here, the major difference from my rationality-first view concerns the order of explanation between principles and virtues as character traits. The idea is that we should understand rationality primarily as a virtue of agents, and then when the agents are fully virtuous, we can read off principles of rationality from the relevant (rational) actions or attitudes of these virtuous agents (Wedgwood, 2017, ch. 6).

However, why could one not just as well argue that principles are prior to virtues, and then understand the virtues of rationally virtuous agents in terms of dispositions that are conducive to making them act on the principles of rationality? I do not see what would

[2] Interestingly, they share this view with sceptics like Kiesewetter and Raz. Note, however, that Lord has started to merge his view with constitutivism in recent work (Lord and Sylvan, 2019).

be lost by taking principles to be prior to virtues, putting rationality first – and as my view also has other attractive upshots, such as being able to explain reasons and moral norms, I do not see why one should not opt for it rather than a virtue-based view.

(4) Conclusion

In section 1, I argued against Kantian, Humean, and other contemporary constitutivists who have presented various types of constitutivism about rationality. In section 2, I argued against sceptics about rationality. In section 3, I argued that the principled rationality-first approach to rationality has explanatory benefits over other leading contenders in the literature.

Instrumental rationality defended? Yes, but I want to finish with a *caveat*. The instrumental principle looks like a very limited conception of rationality. One might want to defend more principles of practical rationality – not to mention principles of theoretical rationality. Indeed, most contemporary writers start off with a number of principles, practical *and* theoretical, and then try to present a view that can make sense of that set of principles (cf. chapter 5, section 1). Maybe my view is insufficient?

There may well be other principles of rationality than the instrumental principle. But, even so, it is likely that the instrumental principle is a building block in a full theory of rationality. I have focused on it because of its potential payoff in a constitutivist views about morality. This explanation is compatible with saying more about other principles.

There are options for how one could do so. Perhaps one way to defend other principles internally to my framework is to hold that they are part of the 'success disjunct' of my action disjunctivism (cf. chapter 5, section 7; appendix A, section 2). Or maybe instrumental rationality is, contrary to appearances, the one principle of (at least practical) rationality – perhaps one can explain other seeming principles away (cf. chapter 5, footnote 15; chapter 6, footnote 16). Or maybe what philosophers have taken rationality to be actually consists of several, not necessarily related, phenomena. There is more to say here, and I hope to do so elsewhere.

References

Alm, D. 2011. Defending Fundamental Requirements of Practical Reason: A Constitutivist Framework. *Journal of Philosophical Research*. 36, pp. 77-102.

Anscombe, G.E.M. 1957. *Intention*. Cambridge, MA: Harvard University Press.

Arpaly, N. and Schroeder, T. 2013. *In Praise of Desire*. Oxford, UK: Oxford University Press.

Arruda, C.T. 2017. Why Care About Being an Agent? *Australasian Journal of Philosophy*. 95(3), pp. 488-504.

Bacharach, M. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. In: Bacharach, M., Gold, N. and Sugden, R. ed. Beyond Individual Choice: Teams and Frames in Game Theory. Princeton, NJ: Princeton University Press.

Bachman, Z. 2018. Moral Rationalism and the Normativity of Constitutive Principles. *Philosophia*. 46(1), pp. 1-19.

Bagnoli, C. ed. 2013. *Constructivism in Ethics*. Cambridge, UK: Cambridge University Press.

Baillie, J. 2000. R*outledge Philosophy Guidebook to Hume on Morality*. London, UK: Routledge.

Barkhausen, M. 2017. Reductionist Moral Realism and the Contingency of Moral Evolution. *Ethics*. 126(3), pp. 662-689.

Bedke, M.S. 2009. Intuitive Non-Naturalism Meets Cosmic Coincidence. *Pacific Philosophical Quarterly*. 90(2), pp. 188-209.

Bedke, M.S. 2010. Rationalist Restrictions and External Reasons. *Philosophical Studies*. 151(1), pp. 39-57.

Benacerraf, P. 1973. Mathematical Truth. *Journal of Philosophy*. 70(19), pp. 661-679.

Berdini, F. Forthcoming. Agency's Constitutive Normativity: An Elucidation. *Journal of Value Inquiry*. [Further details not yet available.]

Björklund, F., Björnsson, G., Eriksson, J., Francén Olinder, R. and Strandberg, C.S. eds. *Motivational Internalism*. Oxford, UK: Oxford University Press.

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford, UK: Oxford University Press.

Boyd, R. 1988. How to Be a Moral Realist. In: Sayre-McCord, G. ed. *Essays on Moral Realism*. Ithaca, NY: Cornell University Press, pp. 191-227.

Brandt, R.B. 1979. *A Theory of the Good and the Right*. Oxford, UK: Oxford University Press.

Bratman, M. 1999. *Intention, Plans and Practical Reason*. Stanford, CA: CSLI Publications.

Bratman, M. 2000. Reflection, Planning, and Temporally Extended Agency. *Philosophical Review*. 109(1), pp. 35-61.

Bratman, M. 2013. *Shared Agency: A Planning Theory of Acting Together*. Oxford, UK: Oxford University Press.

Bratman, M. 2016. *The Rational Dynamics of Planning Agency*. The Pufendorf Lectures. 7-10 June, Lund, SWE. [Accessed 16 June 2019.] Available from: http://www.pufendorf.se/sectione195f.html?id=2864.

Brink, D. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge, UK: Cambridge University Press.

Broome, J. 1999. Normative Requirements. *Ratio*. 12(4), pp. 398-419.

Broome, J. 2007. Wide or Narrow Scope? *Mind*. 116(462), pp. 359-370.

Broome, J. 2008. Replies to Southwood, Kearns and Star, and Cullity. *Ethics*. 119(1), pp. 96-108.

Broome, J. 2013. *Rationality Through Reasoning*. Oxford, UK: Oxford University Press.

Brunero, J. 2017. Recent Work on Internal and External Reasons. *American Philosophical Quarterly*. 54(2), pp. 99-118.

Bukoski, M. 2016. A Critique of Smith's Constitutivism. *Ethics*. 127(1), pp. 116-146.

Calhoun, C. 2009. What Good is Commitment? *Ethics*. 119(4), pp. 613-641.

Chang, R. 2009. Voluntarist Reasons and the Sources of Normativity. In: Sobel, D. and Wall, S. eds. *Reasons for Action*, New York, NY: Cambridge University Press, pp. 243-271.

Chang, R. 2013. Grounding Practical Normativity: Going Hybrid. *Philosophical Studies*. 164(1), pp. 163-187.

Chartier, G. 2018. *The Logic of Commitment*. Oxford, UK: Routledge.

Clark, P. 2001. Velleman's Autonomism. *Ethics*. 111(3), pp. 580-593.

Clarke, R. 2010. Intentional Omissions. In: Aguilar, J., Buckareff, A. and Frankish, K. eds. *New Waves in Philosophy of Action*. London, UK: Palgrave MacMillan, pp. 135-156.

Coons, C. and Faraci, D. 2010. First-Personal Authority and the Normativity of Rationality. *Philosophia*. 38(4), pp. 733-740.

Copp, D. 1995. *Morality, Normativity and Society*. Oxford, UK: Oxford University Press.

Copp, D. 2013. Is Constructivism in Ethics an Alternative to Moral Realism? In: Bagnoli, C. ed. *Constructivism in Ethics*. Cambridge, UK: Cambridge University Press, pp. 108-132.

Dancy, J. 2000. *Practical Reality*. Oxford, UK: Oxford University Press.

Dancy, J. 2018. *Practical Shape: A Theory of Practical Reasoning*. Oxford, UK: Oxford University Press.

Daniels, N. 1996. *Justice and Justification: Reflective Equilibrium in Theory and Practice*. New York, NY: Cambridge University Press.

Davidson, D. 1980*a*. Mental Events. In: Davidson, D. *Essays on Actions and Events*. Oxford, UK: Oxford University Press, pp. 207-225.

Davidson, D. 1980*b*. Actions, Reasons, and Causes. In: Davidson, D. *Essays on Actions and Events*. Oxford, UK: Oxford University Press, pp. 3-19.

Davidson, D. 1980*c*. Hempel on Explaining Action. In: Davidson, D. *Essays on Actions and Events*. Oxford, UK: Oxford University Press, pp. 261-275.

Davidson, D. 1987. Knowing One's Own Mind. *Proceedings and Addresses of the American Philosophical Association***.** 60(3), pp. 441-458.

de Waal, F. 2006. Primates and Philosophers: How Morality Evolved. In: de Waal, F., Macedo, S. and Ober, J. eds. *Primates and Philosophers: How Morality Evolved.* Princeton, NJ: Princeton University Press.

DeMetriou, K. 2010. The Soft-Line Solution to Pereboom's Four-Case Argument. *Australasian Journal of Philosophy*. 88(4), pp. 595-617.

Dick, D.G. 2017. Constitutivism, Error, and Moral Responsibility in Bishop Butler's Ethics. *Southern Journal of Philosophy*. 55(4), pp. 415-438.

Doris, J.M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford, UK: Oxford University Press.

Dorsey, D. 2016. *The Limits of Moral Authority*. Oxford, UK: Oxford University Press.

Dorsey, D. 2018. A Perfectionist Humean Constructivism. *Ethics*. 128(3), pp. 574-602.

Dreier, J. 1997. Humean Doubts about the Practical Justification of Morality. In: Cullity, G. and Gaut, B. eds. 1997. *Ethics and Practical Reason*. Oxford, UK: Oxford University Press, pp. 81-100.

Dreier, J. 2015. Can Reasons Fundamentalism Answer the Normative Question? In: Björklund, F., Björnsson, G., Eriksson, J., Francén Olinder, R. and Strandberg, C.S. eds. *Motivational Internalism*. Oxford, UK: Oxford University Press, pp. 167-181.

Driver, J. 2016. Contingency and Constructivism. In: Kirchin, S. ed. *Reading Parfit: On What Matters*. London, UK: Routledge, pp. 172-188.

Enç, B. 2003. *How We Act: Causes, Reasons, and Intentions*. Oxford, UK: Oxford University Press.

Enoch, D. 2006. Agency, Shmagency: Why Normativity Won't Come from What is Constitutive of Action. *Philosophical Review*. 115(2), pp. 169-198.

Enoch, D. 2011*a*. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford, UK: Oxford University Press.

Enoch, D. 2011*b*. Shmagency Revisited. In: Brady, M. ed. *New Waves in Metaethics*. New York, NY: Palgrave MacMillan, pp. 208-233.

Ferrero, L. 2009. Constitutivism and the Schmagency Challenge. In: Shafer-Landau, R. ed. *Oxford Studies in Metaethics*. 4, pp. 303-332.

Ferrero, L. 2012. Diachronic Constraints on Practical Rationality. *Philosophical Issues*. 22(1), pp. 144-164.

Ferrero, L. 2015. Katsafanas, Paul. Agency and the Foundations of Ethics: Nietzschean Constitutivism. *Ethics*. 125(3), pp. 883-888.

Ferrero, L. 2018. Inescapability Revisited. *Manuscrito*. 41(4), pp. 113-158.

Ferrero, L. 2019. The Simple Constitutivist Move. *Philosophical Explorations*. 22(2), pp. 146-162.

Fine, K. 2002. Varieties of Necessity. In: Szabo Gendler, T. and Hawthorne, J. eds. *Conceivability and Possibility*. Oxford, UK: Oxford University Press, pp. 253-281.

Fink, J. 2014. A Constitutive Account of 'Rationality Requires'. *Erkenntnis*. 79(4), pp. 1-33.

Finlay, S. 2019. Defining Normativity. In: Plunkett, D., Shapiro, S.J. and Toh, K. eds. *Dimensions of Normativity: New Essays on Metaethics and Jurisprudence*. Oxford, UK: Oxford University Press, pp. 187-220.

Finlay, S. and Schroeder, M. 2017. Reasons for Action: Internal vs. External. In: Zalta, E.N. ed. *The Stanford Encyclopedia of Philosophy*. Fall 2017 edition. [Online.] [Accessed 6 June 2019.] Available from: https://plato.stanford.edu/archives/fall2017/entries/reasons-internal-external/.

Fischer, J.M. 2004. Responsibility and Manipulation. *Journal of Ethics*. 8(2), pp. 145-177.

Foot, P. 1972. Morality as a System of Hypothetical Imperatives. *Philosophical Review*. 81(3), pp. 305-316.

Foot, P. 2001. *Natural Goodness*. Oxford, UK: Oxford University Press.

Forcehimes, A. and Semrau, L. 2018. Are There Distinctively Moral Reasons? *Ethical Theory and Moral Practice*. 21(3), pp. 699-717.

Frankfurt, H. 1969. Alternate Possibilities and Moral Responsibility. *Journal of Philosophy*. 66(23), pp. 829-839.

Frankfurt, H. 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy*. 68(1), pp. 5-20.

Frey, J. 2019. Happiness as the Constitutive Principle of Action in Thomas Aquinas. *Philosophical Explorations*. 22(2), pp. 208-221.

Gauthier, D. 1985. *Morals by Agreement*. Oxford, UK: Oxford University Press.

Gauthier, D. 2013. Twenty-Five On. *Ethics*. 123(4), pp. 601-624.

Gibbard, A. 1999. Morality as Consistency in Living: Korsgaard's Kantian Lectures. *Ethics*. 110(1), pp. 140-164.

Gibbard, A. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Goldman, A. 2011. *Reasons from Within: Desires and Values*. Oxford, UK: Oxford University Press.

Goldman, A.I. 1979. What Is Justified Belief? In: Pappas, G.S. ed. *Justification and Knowledge*. Dordrecht, NE: D. Reidel, pp. 1-25.

Hampton, J. 1998. *The Authority of Reason*. Cambridge, UK: Cambridge University Press.

Hare, R.M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford, UK: Oxford University Press.

Haslanger, S. 2018. *Ameliorating Social Practices: Cognition, Culture, and Critique*. The Mangoletsi Lectures, University of Leeds. 15-23 May, Leeds, UK.

Hazlett, A. 2009. Review of J. David Velleman, How We Get Along. *Notre Dame Philosophical Reviews*. [Online]. [Accessed 14 September 2013]. Available from: http://ndpr.nd.edu/news/24222/?id=18085.

Heathwood, C. 2011. Desire-Based Theories of Reasons, Pleasure, and Welfare. In: Shafer-Landau, R. *Oxford Studies in Metaethics*. 6. Oxford, UK: Oxford University Press, pp. 79-106.

Hedden, B. 2015. *Reasons Without Persons: Rationality, Identity, and Time*. Oxford, UK: Oxford University Press.

Hempel, C. 1961. Rational Action. *Proceedings and Addresses of the American Philosophical Association*. 35, pp. 5-23.

Henning, T. 2018. *From A Rational Point of View. How We Represent Subjective Perspectives in Practical Discourse*. Oxford, UK: Oxford University Press.

Heuer, U. 2001. *Gründe und Motive: Über Humesche Theorien Praktische Vernunft*. Paderborn, GER: Menthis.

Heuer, U. 2004. Reasons for Actions and Desires. *Philosophical Studies*. 121(4), pp. 43-63.

Horgan, T. and Timmons, M. 1991. New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research*. 16(4), pp. 447-465.

Hornsby, J. 2004. Agency and Actions. *Royal Institute of Philosophy Supplement*. 55, pp. 1-23.

Hornsby, J. 2008. A Disjunctive Conception of Acting for Reasons. In: Haddock, A. and McPherson, F. eds. *Disjunctivism: Perception, Action, Knowledge*. Oxford, UK: Oxford University Press, pp. 244-261.

Hubin, D. 2001. The Groundless Normativity of Instrumental Rationality. *Journal of Philosophy*. 98(9), pp. 445-468.

Huemer, M. 2005. *Ethical Intuitionism*. New York, NY: Palgrave MacMillan.

Hume, D. 1978. A Treatise of Human Nature. In: Selby-Bigge, L.A. and Nidditch, P.H. eds. *A Treatise of Human Nature*. 2nd edition. Oxford, UK: Oxford University Press.

Hurley, S. 2001. Reason and Motivation: The Wrong Distinction? *Analysis*. 61(2), pp. 151-155.

Hursthouse, R. 1991. Arational Actions. *Journal of Philosophy*. 88(2), pp. 57-68.

Hurtig, K.I. 2006. Internalism and Accidie. *Philosophical Studies*. 129(3), pp. 517-543.

Hussain, N.J.Z. ms. The Requirements of Rationality. [Online.] [Accessed 6 June 2019.] Available from: https://web.stanford.edu/~hussainn/StanfordPersonal/Online_Papers_files/HussainRequirementsv24.pdf.

Hussain, N.J.Z. and Shah, N. 2006. Misunderstanding Metaethics: Korsgaard's Rejection of Realism. In: Shafer-Landau, R. ed. *Oxford Studies in Metaethics*. 1. Oxford, UK: Oxford University Press, pp. 265-294.

Hussain, N.J.Z. and Shah, N. 2013. Meta-ethics and its Discontents: A Case Study of Korsgaard. In: Bagnoli, C. ed. 2013. *Constructivism in Ethics*. Cambridge, UK: Cambridge University Press, pp. 82-107.

Johnson, R. N. 1997. Reasons and Advice for the Practically Rational. *Philosophy and Phenomenological Research*. 57(3), pp. 619-625.

Johnson, R. N. 1999. Internal Reasons and the Conditional Fallacy. *Philosophical Quarterly*. 49(194), pp. 53-72.

Johnson, R. N. 2003. Internal Reasons: Reply to Brady, van Roojen and Gert. *Philosophical Quarterly*. 53(213), pp. 573-580.

Joyce, R. 2001. *The Myth of Morality*. Oxford, UK: Oxford University Press.

Joyce, R. 2006. *The Evolution of Morality*. Cambridge, MA: MIT Press.

Kant, I. 1996. Groundwork for the Metaphysics of Morals. In: Gregor, M.J. ed. *The Cambridge Edition of the Works of Immanuel Kant – Practical Philosophy*. Cambridge, UK: Cambridge University Press, pp. 37-108.

Katsafanas, P. 2013. *Agency and the Foundations of Ethics: Nietzschean Constitutivism*. Oxford, UK: Oxford University Press.

Katsafanas, P. 2015. Value, Affect, and Drive. In: Kail, P.J.E. and Dries, M. eds. *Nietzsche on Mind and Nature*. Oxford, UK: Oxford University Press, pp. 163-188.

Katsafanas, P. 2018. Constitutivism about Practical Reasons. In: Star, D. ed. *Oxford Handbook of Reasons and Normativity*. Oxford, UK: Oxford University Press, pp. 367-391.

Kauppinen, Antti. Forthcoming. Rationality as the Rule of Reason. *Noûs*. [Further details not yet available.] [Online.] [Accessed 26 September 2019.] Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12311.

Kiesewetter, B. 2017. *The Normativity of Rationality*. Oxford, UK: Oxford University Press.

Kolnai, A. 1961. Deliberation Is of Ends. *Proceedings of the Aristotelian Society*. 62(1), pp. 195-218.

Kolodny, N. 2005. Why Be Rational? *Mind*. 114(455), pp. 509-563.

Kolodny, N. 2007*a*. How Does Coherence Matter? *Proceedings of the Aristotelian Society*. 107(1pt3), pp. 229-263.

Kolodny, N. 2007*b*. State or Process Requirements? *Mind*. 116(462), pp. 371-385.

Kolodny, N. 2008*a*. The Myth of Practical Consistency. *European Journal of Philosophy*. 16(3), pp. 366-402.

Kolodny, N. 2008*b*. Why Be Disposed to be Coherent? *Ethics*. 118(3), pp. 437-463.

Kolodny, N. 2009. Reply to Bridges. *Mind*. 118(470), pp. 369-376.

Korsgaard, C.M. 1986. Skepticism about Practical Reason. *Journal of Philosophy*. 83(1), pp. 1-25.

Korsgaard, C.M. 1996*a*. *The Sources of Normativity*. Cambridge, UK: Cambridge University Press.

Korsgaard, C.M. 1996*b*. Morality as Freedom. In: Korsgaard, C.M. *Creating the Kingdom of Ends*. Cambridge, UK: Cambridge University Press, pp. 159-187.

Korsgaard, C.M. 1997. The Normativity of Instrumental Reason. In: Cullity, G. and Gaut, B. eds. 1997. *Ethics and Practical Reason*. Oxford, UK: Oxford University Press, pp. 215-254.

Korsgaard, C.M. 1999. Self-Constitution in the Ethics of Plato and Kant. *Journal of Ethics*. 3(1), pp. 1-29.

Korsgaard, C.M. 2003. Realism and Constructivism in Twentieth Century Moral Philosophy. *Journal of Philosophical Research (Supplement)*. 28, pp. 99-122.

Korsgaard, C.M. 2005. Acting for a Reason. *Danish Yearbook of Philosophy*. 40, pp. 11-36.

Korsgaard, C.M. 2008. *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford, UK: Oxford University Press.

Korsgaard, C.M. 2009. *Self-Constitution: Agency, Identity, Integrity*. Oxford, UK: Oxford University Press.

Korsgaard, C.M. 2014. The Normative Constitution of Agency. In: Vargas, M. and Yaffe, G. eds. *Rational and Social Agency: The Philosophy of Michael Bratman*. Oxford, UK: Oxford University Press, pp. 190-214.

Korsgaard, C.M. 2018. *Fellow Creatures: Our Obligations to the Other Animals*. Oxford, UK: Oxford University Press.

Korsgaard, C.M. 2019. Constitutivism and the Virtues. *Philosophical Explorations*. 22(2), pp. 98-116.

Korsgaard, C.M. and Pauer-Studer, H. 2002. *Internalism and the Sources of Normativity*. [Online.] [Accessed 16 June 2019.] Available from: http://www.people.fas.harvard.edu/~korsgaar/CPR.CMK.Interview.pdf.

Kosch, M. 2018. *Fichte's Ethics*. Oxford, UK: Oxford University Press.

Lavin, D. 2004. Practical Reason and the Possibility of Error. *Ethics*. 114(3), pp. 424-457.

Lavin, D. 2017. Forms of Rational Agency. *Royal Institute of Philosophy Supplement*. 80, pp. 171-193.

Leffler, O. 2014. *Rationalism with a Humean Face*. MA Dissertation, University of Gothenburg, SWE.

Leffler, O. 2016. The Foundations of Agency – and Ethics? *Philosophia*. 44(2), pp. 547-563.

Leffler, O. 2019. New Shmagency Worries. *Journal of Ethics and Social Philosophy*. 15(2), pp. 121-145.

Lenman, J. 2009. The Politics of the Self: Stability, Normativity and the Lives We Can Live with Living. In: Bortolotti, L. ed. *Philosophy and Happiness*. London, UK: Palgrave MacMillan, pp. 183-199.

Lenman, J. and Shemmer, Y. eds. 2012. *Constructivism in Practical Philosophy*. Oxford, UK: Oxford University Press.

Levy, Y. 2018. Does the Normative Question about Rationality Rest on a Mistake? *Synthese*. 195(5), pp. 2021-2038.

Lindeman, K. 2017. Constitutivism without Normative Thresholds. *Journal of Ethics and Social Philosophy*. 12(3), pp. 231-258.

Lindeman, K. 2019. Functional Constitutivism's Misunderstood Resources: A Limited Defense of Smith's Constitutivism. *Ethics*. 130(1), pp. 79-91.

Lindenberg, S. and Steg, L. 2007. Normative, Gain and Hedonic Goal Frames Guiding Environmental Behavior. *Journal of Social Issues*. 63(1), pp. 117-137.

Lord, E. 2018. *The Importance of Being Rational*. Oxford, UK: Oxford University Press.

Lord, E. and Sylvan, K. 2019. Reasons: Wrong, Right, Normative, Fundamental. In: *Journal of Ethics and Social Philosophy*. 15(1), pp. 43-74.

Lynch, M.P. 2001. *The Nature of Truth: Classic and Contemporary Readings*. Cambridge, MA: MIT Press.

MacIntyre, A. 2016. *Ethics in the Conflicts of Modernity: An Essay on Desire, Practical Reasoning, and Narrative*. Cambridge, UK: Cambridge University Press.

Mackie, J.L. 1977. *Ethics: Inventing Right and Wrong*. New York, US: Viking Press.

Manne, K. 2014. Internalism about Reasons: Sad but True? *Philosophical Studies*. 167(1), pp. 89-117.

Manne, K. 2016. Democratizing Humeanism. In: Lord, E. and Maguire, B. eds. *Weighing Reasons*. Oxford, UK: Oxford University Press, pp. 123-140.

Markovits, J. 2014. *Moral Reason*. Oxford, UK: Oxford University Press.

Martin, M.G.F. 2004. The Limits of Self-Awareness. *Philosophical Studies*. 120(1-3), pp. 37-89.

Mayr, E. 2011. *Understanding Human Agency*. Oxford, UK: Oxford University Press.

Mayr, E. 2019. Blame for Constitutivists: Kantian Constitutivism and the Victim's Special Standing to Complain. *Philosophical Explorations*. 22(2), pp. 117-129.

McClennen, E.F. 2004. The Rationality of Being Guided by Rules. In: Mele, A.R. and Rawling, P. eds. *The Oxford Handbook of Rationality*. Oxford, UK: Oxford University Press, pp. 222-239.

McKenna, M. 2008. A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument. *Philosophy and Phenomenological Research*. 77(1), pp. 142-159.

McPherson, T. 2015. Supervenience in Ethics. In: Zalta, E.N. ed. *The Stanford Encyclopedia of Philosophy*. Winter 2015 edition. [Online.] [Accessed 11 September 2019.] Available from: https://plato.stanford.edu/archives/win2015/entries/supervenience-ethics/.

Melden, A.I. 1961. *Free Action*. Oxford, UK: Routledge.

Mele, A.R. 2003. *Motivation and Agency*. Oxford, UK: Oxford University Press.

Millgram, E. 1996. Williams' Argument Against External Reasons. *Noûs*. 30(2), pp. 197-220.

Millgram, E. 2011. Critical Notice: Self-Constitution: Agency, Identity, and Integrity / The Constitution of Agency: Essays on Practical Reason and Moral Psychology. *Australasian Journal of Philosophy*. 89(3), pp. 549-556.

Moore, G.E. 1903. *Principia Ethica*. Amherst, NY: Prometheus Books.

Mylonaki, E. 2018. Instrumental Normativity and the Practicable Good: A Murdochian Constitutivist Account. *Manuscrito*. 41(4), pp. 349-388.

Nagel, T. 1970. *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

Nagel, T. 1986. *The View from Nowhere*. Oxford, UK: Oxford University Press.

Narveson, J. 2001. *The Libertarian Idea*. Peterborough, CA: Broadview Press.

North, D.C. 1991. Institutions. *Journal of Economic Perspectives*. 5(1), pp. 97-112.

O'Brien, L. 2018. Action Explanation and Its Presuppositions. *Canadian Journal of Philosophy*. 49(1), pp. 123-146.

Oddie, G. 2005. *Value, Reality, and Desire*. Oxford, UK: Oxford University Press.

O'Hagan, E. 2014. Shmagents, Realism and Constitutivism about Rational Norms. *Journal of Value Inquiry*. 48(1), pp. 17-31.

Olson, J. 2014. *Moral Error Theory: History, Critique, Defence*. Oxford, UK: Oxford University Press.

Olsson, E.J. 2005. *Against Coherence. Truth, Probability, and Justification*. Oxford, UK: Oxford University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Paakkunainen, H. 2017. Normativity and Agency. In: McPherson, T. and Plunkett, D. eds. *The Routledge Handbook of Metaethics*. New York, NY: Routledge, pp. 402-416.

Paakkunainen, H. 2018. Doing Away with the 'Shmagency' Objection to Constitutivism. *Manuscrito*. 41(4), pp. 431-80.

Parfit, D. 1984. *Reasons and Persons*. Oxford, UK: Oxford University Press.

Parfit, D. 2011*a*. *On What Matters*. Volume 1. Oxford, UK: Oxford University Press.

Parfit, D. 2011*b*. *On What Matters*. Volume 2. Oxford, UK: Oxford University Press.

Pauer-Studer, H. 2018. Korsgaard's Constitutivism and the Possibility of Bad Action. *Ethical Theory and Moral Practice*. 21(1), pp. 37-56.

Pereboom, D. 2001. *Living Without Free Will*. Cambridge, UK: Cambridge University Press.

Persson, I. 2013. *From Morality to the End of Reason*. Oxford, UK: Oxford University Press.

Peter, F. 2019. Normative Facts and Reasons. *Proceedings of the Aristotelian Society*. 119(1), pp. 53-75.

Petersson, B. 2000. *Belief & Desire: The Standard Model of Intentional Action – Critique and Defence*. PhD Dissertation, University of Lund, SWE.

Pettit, P. and Smith, M. 1990. Backgrounding Desire. *Philosophical Review*. 99(4), pp. 565-592.

Railton, P. 1984. Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs*, 13(2), pp. 134-171.

Railton, P. 1986. Moral Realism. *Philosophical Review*. 95(2), pp. 163-207.

Railton, P. 1997. On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action. In: Cullity, G. and Gaut, B. eds. 1997. *Ethics and Practical Reason*. Oxford, UK: Oxford University Press, pp. 53-80.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 1993. *Political Liberalism*. New York, NY: Columbia University Press.

Raz, J. 2005*a*. The Myth of Instrumental Rationality. *Journal of Ethics and Social Philosophy*. 1(1), pp. 1-28.

Raz, J. 2005*b*. Instrumental Rationality: A Reprise. *Journal of Ethics and Social Philosophy*. 1(1), pp. 1-20.

Ridge, M. 2014. *Impassioned Belief*. Oxford, UK: Blackwell.

Ridge, M. 2018. Meeting Constitutivists Halfway. *Philosophical Studies*. 175(12), pp. 2951-2968.

Regan, D. 1980. *Utilitarianism and Cooperation*. Oxford, UK: Clarendon Press.

Rosati, C.S. 2016. Agents and Shmagents: An Essay on Agency and Normativity. In: Shafer-Landau, R. ed. *Oxford Studies in Metaethics*. 11. Oxford, UK: Oxford University Press, pp. 182–213.

Ruben, D-H. 2003. *Action and Its Explanation*. Oxford, UK: Oxford University Press.

Scanlon, T.M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T.M. 2007. Structural Irrationality. In: Brennan, G., Goodin, R., Jackson, F. and Smith, M. eds. *Common Minds: Themes from the Philosophy of Philip Pettit*. Oxford, UK: Oxford University Press, pp. 84-103.

Scanlon, T.M. 2014. *Being Realistic about Reasons*. Oxford, UK: Oxford University Press.

Schaab, J.D. 2019. *Kantian Constructivism: A Restatement*. PhD Dissertation. University of St Andrews, UK.

Schapiro, T. 2001. Three Conceptions of Action in Moral Theory. *Noûs*. 35(1), pp. 32-57.

Schlosser, M.E. 2010. Agency, Ownership, and the Standard Theory. In: Aguilar, J., Buckareff, A. and Frankish, K. eds. *New Waves in Philosophy of Action*. London, UK: Palgrave MacMillan, pp. 13-31.

Schroeder, M. 2007. *Slaves of the Passions*. Oxford, UK: Oxford University Press.

Schroeder, M. 2011. Ought, Agents, and Actions. *Philosophical Review*. 120(1), pp. 1-41.

Schroeder, M. 2014. *Explaining the Reasons We Share: Explanation and Expression in Ethics*. Volume 1. Oxford, UK: Oxford University Press.

Schroeder, T. 2015. Desire. In: Zalta, E.N. ed. *The Stanford Encyclopedia of Philosophy*. Summer 2017 edition. [Online.] [Accessed 6 June 2019.] Available from: https://plato.stanford.edu/archives/sum2017/entries/desire/.

Schueler, G.F. 2009. The Humean Theory of Motivation Rejected. *Philosophy and Phenomenological Research*. 78(1), pp. 103-122.

Setiya, K. 2007. *Reasons Without Rationalism*. Princeton, NJ: Princeton University Press.

Shafer-Landau, R. 2003. *Moral Realism: A Defence*. Oxford, UK: Oxford University Press.

Shope, R.K. 1978. The Conditional Fallacy in Contemporary Philosophy. *Journal of Philosophy*. 75(8), pp. 397-413.

Silverstein, M. 2015. The Shmagency Question. *Philosophical Studies*. 172(5), pp. 1127-1142.

Silverstein, M. 2016. Teleology and Normativity. In: Shafer-Landau, R. ed. *Oxford Studies in Metaethics*. 11. Oxford, UK: Oxford University Press, pp. 214-240.

Sinhababu, N. 2009. The Humean Theory of Motivation Reformulated and Defended. *Philosophical Review*. 118(4), pp. 465-500.

Sinhababu, N. 2011. The Humean Theory of Practical Irrationality. *Journal of Ethics and Social Philosophy*. 6(1), pp. 1-13.

Sinhababu, N. 2017. *Humean Nature: How Desire Explains Action, Thought, and Feeling*. Oxford, UK: Oxford University Press.

Skorupski, J. 2010. *The Domain of Reasons*. Oxford, UK: Oxford University Press.

Skow, B. 2016. *Reasons Why*. Oxford, UK: Oxford University Press.

Smith, M. 1994. *The Moral Problem*, Oxford, UK: Blackwell Publishing.

Smith, M. 1995. Internal Reasons. *Philosophy and Phenomenological Research*. 55(1), pp. 109-131.

Smith, M. 1999. Search for the Source. *Philosophical Quarterly*. 49(169), pp. 384-394.

Smith, M. 2003*a*. Rational Capacities, Or: How to Distinguish Recklessness, Weakness, and Compulsion. In: Stroud, S. and Tappolet, C. eds. *Weakness of Will and Practical Irrationality*. Oxford, UK: Clarendon Press, pp. 17-38.

Smith, M. 2003*b*. Humeanism, Psychologism, and the Normative Story. *Philosophy and Phenomenological Research*. 67(2), pp. 460-467.

Smith, M. 2009. The Explanatory Role of Being Rational. In: Sobel, D. and Wall, S. eds. *Reasons for Action*. New York, NY: Cambridge University Press, pp. 58-80.

Smith, M. 2011. Deontological Moral Obligations and Non-Welfarist Agent-Relative Values. *Ratio*. 24(4), pp. 351-363.

Smith, M. 2012*a*. Agents and Patients: Or, What We Learn about Reasons for Action by Reflecting on Our Choices in Process-of-Thought Cases. *Proceedings of the Aristotelian Society*. 112(3), pp. 309-331.

Smith, M. 2012*b*. Four Objections to the Standard Story of Action (and Four Replies). *Philosophical Issues*. 22(1), pp. 387-401.

Smith, M. 2012*c*. A Puzzle about Internal Reasons. In Heuer, U. and Lang, G. eds. *Luck, Value and Commitment: Themes from the Philosophy of Bernard Williams*. Oxford, UK: Oxford University Press, pp. 195-218.

Smith, M. 2013. A Constitutivist Theory of Reasons: Its Promise and Parts. *LEAP: Law, Ethics, and Philosophy*. 1, pp. 9-30.

Smith, M. 2015. The Magic of Constitutivism. *American Philosophical Quarterly*. 52(2), pp. 187-200.

Smith, M. 2017. Constitutivism. In: McPherson, T. and Plunkett, D. eds. *The Routledge Handbook of Metaethics*. New York, NY: Routledge, pp. 371-84.

Smith, M. 2018. *From a Constitutivist Account of Moral Reasons to a Comprehensive Moral Doctrine*. Future of Normativity, University of Kent. 7 June, Canterbury, UK.

Sosa, E. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Volume I. Oxford, UK: Oxford University Press.

Sosa, E. 2009. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Volume II. Oxford, UK: Oxford University Press.

Southwood, N. 2008. Vindicating the Normativity of Rationality. *Ethics*. 119(1), pp. 9-30.

Southwood, N. 2018*a*. Constructivism about Reasons. In: Star, D. ed. *Oxford Handbook of Reasons and Normativity*. Oxford, UK: Oxford University Press, pp. 342-366.

Southwood, N. 2018*b*. Constructivism and the Normativity of Practical Reason. In: Jones, K. and Schroeter, F. eds. *The Many Moral Rationalisms*. Oxford, UK: Oxford University Press, pp. 91-109.

Stampe, D.W. 1987. The Authority of Desire. *Philosophical Review*. 96(3), pp. 335-381.

Star, D. 2016. *Knowing Better*. Oxford, UK: Oxford University Press.

Steward, H. 2012. Actions as Processes. *Philosophical Perspectives*. 26(1), pp. 373-388.

Stocker, M. 1979. Desiring the Bad: An Essay in Moral Psychology. *Journal of Philosophy*. 76(12), pp. 738-753.

Strandberg, C.S. 2018. Towards an Ecumenical Theory of Normative Reasons. *Dialectica*. 72(1), pp. 69-100.

Strandberg, C.S. 2019. An Ecumenical Account of Categorical Moral Reasons. *Journal of Moral Philosophy*. 16(2), pp. 160-188.

Street, S. 2006. A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*. 127(1), pp. 109-166.

Street, S. 2008. Constructivism about Reasons. In: Shafer-Landau, R. ed. *Oxford Studies in Metaethics*. 3, pp. 207-245.

Street, S. 2009. In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters. *Philosophical Issues*. 19(1), pp. 273-298.

Street, S. 2010. What is Constructivism in Ethics and Metaethics? *Philosophy Compass*. 5(5), pp. 363-384.

Street, S. 2012. Coming to Terms with Contingency. In: Lenman, J. and Shemmer, Y. eds. *Constructivism in Practical Philosophy*. Oxford, UK: Oxford University Press, pp. 40-59.

Svavarsdóttir, S. 2008. The Virtue of Practical Rationality. *Philosophy and Phenomenological Research*. 77(1), pp. 1-33.

Taylor, C. 1985. What is Human Agency? In: Taylor, C. *Human Agency and Language: Philosophical Papers 1*. Cambridge, UK: Cambridge University Press, pp. 15-44.

Tenenbaum, S. 2019. Formalism and Constitutivism in Kantian Practical Philosophy. *Philosophical Explorations*. 22(2), pp. 163-176.

Tersman, F. 1992. Coherence and Disagreement. *Philosophical Studies*. 65(3), pp. 305-317.

Thoma, J. ms. Folk Psychology and the Interpretation of Decision Theory. [Online.] [Accessed 16 June 2019.] Available from: https://johannathoma.files.wordpress.com/2019/05/folk-psychology-and-the-interpretation-of-decision-theory-anonymised0419.pdf.

Thomson, J.J. 2008. *Normativity*. Peru, IL: Open Court Publishing.

Tiffany, E. 2011. Why Be an Agent? *Australasian Journal of Philosophy*. 90(2), pp. 223-233.

Tubert, A. 2010. Constitutive Arguments. *Philosophy Compass*. 5(8), pp. 656-666.

Velleman, J.D. 1992. What Happens When Someone Acts? *Mind*. 101(403), pp. 461-481.

Velleman, J.D. 1996. The Possibility of Practical Reason. *Ethics*. 106(4), pp. 694-726.

Velleman, J.D. 1997. Deciding How to Decide. In: Cullity, G. and Gaut, B. eds. *Ethics and Practical Reason*. Oxford, UK: Oxford University Press, pp. 29-52.

Velleman, J.D. 2000. *The Possibility of Practical Reason*. Oxford, UK: Oxford University Press.

Velleman, J.D. 2006. *Self to Self: Selected Essays*. New York, NY: Cambridge University Press.

Velleman, J.D. 2007*a*. What Good is a Will? In: Leist, A. ed. *Action in Context*. De Gruyter, GER: pp. 193-215.

Velleman, J.D. 2007*b*. *Practical Reflection*. Stanford, CA: CSLI Publications.

Velleman, J.D. 2009. *How We Get Along*. Oxford, UK: Oxford University Press.

Velleman, J.D. 2013. *Foundations for Moral Relativism*. Cambridge, UK: Open Book Publishing.

Velleman, J.D. and Shah, N. 2005. Doxastic Deliberation. *Philosophical Review*. 114(4), pp. 497-534.

Väyrynen, P. 2013. Grounding and Normative Explanation. *Aristotelian Society Supplementary Volume*. 87(1), pp. 155-187.

Walden, K. 2012. Laws of Nature, Laws of Freedom, and the Social Construction of Normativity. In: Shafer-Landau, R. ed. 2012. *Oxford Studies in Metaethics*. 7, pp. 37-79.

Walden, K. 2018. Practical Reason Not as Such. *Journal of Ethics and Social Philosophy*. 13(2), pp. 125-153.

Wallace, R.J. 2001. Normativity, Commitment, and Instrumental Reason. *Philosophers' Imprint*. 1, pp. 1-26.

Wallace, R.J. 2003. Review of Richard Joyce, The Myth of Morality. *Notre Dame Philosophical Reviews*. [Online]. [Accessed 16 June 2019]. Available from: https://ndpr.nd.edu/news/the-myth-of-morality/.

Wallace, R.J. 2004. Normativity and the Will. *Royal Institute of Philosophy Supplement*. 55, pp. 195-216.

Wallace, R.J. 2012. Constructivism about Normativity: Some Pitfalls. In: Lenman, J. and Shemmer, Y. eds. 2012. *Constructivism in Practical Philosophy*. Oxford, UK: Oxford University Press, pp. 18-39.

Wallace, R.J. 2019. *The Moral Nexus*. Princeton, NJ: Princeton University Press.

Wedgwood, R. 2017. *The Value of Rationality*. Oxford, UK: Oxford University Press.

Wiland, E. 2000. Good Advice and Rational Action. *Philosophy and Phenomenological Research*. 60(3), pp. 561-569.

Wiland, E. 2003. Some Advice for Moral Psychologists. *Pacific Philosophical Quarterly*. 84(3), pp. 299-310.

Wiland, E. 2012. *Reasons*. London, UK: Continuum.

Williams, B.A.O. 1981. Internal and External Reasons. In: Williams, B.A.O. *Moral Luck: Philosophical Papers 1973-1980*. Cambridge, UK: Cambridge University Press, pp. 101-13.

Williams, B.A.O. 1985. *Ethics and the Limits of Philosophy*. Oxford, UK: Oxford University Press.

Williams, B.A.O. 1995. Internal Reasons and the Obscurity of Blame. In: Williams, B.A.O. *Making Sense of Humanity and Other Philosophical Papers 1982-1993*. Cambridge, UK: Cambridge University Press.

Williamson, T. 2001. *Knowledge and Its Limits*. Oxford, UK: Oxford University Press.

Wittgenstein, L. 2009. *Philosophical Investigations*. In: Anscombe, G.E.M. trans., Hacker, P.M.S. and Schulte, J. eds. Philosophical Investigations. 4th edition. Chichester and Oxford, UK: Wiley-Blackwell.