# A constitutivist theory of reasons: its promise and parts [*]

MICHAEL SMITH

**Abstract**

The aim of this paper is two-fold. First, it explains what a constitutivist theory of reasons is and why the theory promises to deliver the holy grail of moral philosophy, which is an argument to the conclusion that each of us would choose to act morally if we had and exercised the capacity to respond rationally to the circumstances in which we find ourselves. Second, it describes the various parts of a constitutivist theory of reasons, and it explains how these parts give support to the premises required for the promised argument.

**Keywords**: Reasons, constitutivist theory, moral action, rationality.

## 1. PROMISE

Philosophers have long felt the need to provide morality with a solid foundation. Among the ways in which they have tried to do this, perhaps the most common, the most optimistic, and usually the most disappointing, has been the attempt to reduce facts about the wrongness of actions to facts about the rational requirements to which they are subject (see for example Kant 1786, Nagel 1970, and Korsgaard 1996, but contrast Hume 1740, and Blackburn 1998). According to the rationalist, wrong acts are irrational. Nor should this be surprising given the nature of action on the one hand, and the reason why morality's foundations have seemed to be in such dire need

[*] An earlier version of this paper was delivered as the Inaugural LEAP Lecture, Universitat Pompeu Fabra, Barcelona, in 2011. Special thanks are due to Andrew Williams, my commentator on that occasion, and to the members of the audience for their many helpful, if rather skeptical, comments and questions. The paper was written while I was visiting the Humboldt University in Berlin, Germany, on a Humboldt Research Award provided by the Alexander von Humboldt Foundation. I am very grateful to the Foundation for its support. The paper draws on, and attempts to provide the overarching rationale for, two others written during this period: "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values" in *Ratio* (24) 2011 pp.351-363, and "Agents and Patients, Or: What We Learn about Reasons for Action by Reflecting on Process-of-Thought Cases" forthcoming in *Proceedings of the Aristotelian Society*, 2012.

shoring up on the other. Anti-rationalism, it turns out, is very difficult to avoid. The promise of a Constitutivist Theory of Reasons is that it can do the required shoring up, notwithstanding the formidable difficulties involved in providing such a rationalist reduction (see for example Korsgaard 1996, Velleman 2005, but contrast Enoch 2006, and Enoch 2011).

As regards the nature of action, the key point to remember is the pervasive influence of the standard story of action we have inherited from Hume (Hume 1740, see also Davidson 1963, Davidson 1971, and Smith 1998). What makes an agent's movement of her body an action, according to this standard story, is the fact that that movement is produced in the right kind of way by two psychological states: some background final desire the agent has —that is, some desire she has for something for its own sake— and some belief she has about how her bodily movement will bring about, or perhaps constitute, the realization of that final desire. According to the most plausible version of this standard story, these two states combine, under the influence of the agent's exercise of her capacity to be instrumentally coherent, so as to constitute an instrumental desire whose immediate causal role is to move her body (Hempel 1961, Smith 2012).

The main attraction of this story lies in its economy. The concept of an action is no longer a fundamental concept, but is rather analyzable in terms of psychological concepts in which we have to traffic anyway. A further attraction is that is that it makes various other concepts, concepts that we might otherwise find mysterious, much more transparent. Let me give two such examples.

Since it is in the nature of beliefs and desires to be subject to rational requirements, and since actions are analyzable in terms of their relations to beliefs and desires, the standard story of action tells us that the rational standing of actions can be defined in terms of the rational standing of the beliefs and desires that produce them. The rational evaluation of an agents' actions thus requires us to ask questions like these: Are the beliefs that produce agents' actions —that is, beliefs about the likelihood of their bodily movements bringing about, or constituting, their getting what they finally desire— well-supported by the evidence available to them? Are the strengths of their instrumental desires proportional to the strength of their background final desires and the probabilities that they assign to the objects of their instrumental desires being ways of bringing about what they finally desire? In other words, are the agents instrumentally coherent?

To the extent that questions like these get negative answers, the beliefs and desires that produce actions violate the rational requirements to which they are subject —the norm of total evidence on belief and instrumental coherence on instrumental desire— and the actions that those beliefs and desires produce therefore violate rational requirements as well. The actions are irrational in virtue of the irrationality of the beliefs and instrumental desires that produce them. To the extent that these questions get positive answers, the beliefs and instrumental desires that produce actions, and the actions that they produce, conform to rational requirements (see also Smith 2004). So, at any rate, the standard story suggests.

Here is a second example, consequent upon the first. There is a long-standing metaphysical puzzle about the freedom required for the allocation of responsibility (Pettit and Smith 1996). Though this has led some to posit a faculty of choice, immediately prior to action, the standard story of action suggests that this is a gratuitous posit. Think again about the different ways in which we can explain why people's beliefs and desires violate some rational requirement. People might *have but not exercise* the capacity to access the evidence available to them, or the capacity to believe according to the evidence they access, or the capacity to form instrumental desires whose strength is proportional to their background beliefs about the likelihood that the means will achieve what they finally desire and the strength of their final desires. Alternatively, people might *lack* one or another of these capacities. These two explanations in turn have an analytic tie to fault, and hence to responsibility.

If people possess, but fail to exercise the relevant capacities, then there is something that they could have done in a perfectly mundane sense: they could have exercised the capacity to access the available evidence, or to believe in accordance with their evidence, or to instrumentally desire, in the way that's rationally required. Their failure to access or believe or instrumentally desire correctly is therefore their fault, in that same perfectly mundane sense, and so too is their failure to perform the act that they would otherwise have performed. But if people lack one or another of these capacities, then their failure to access or to believe or to instrumentally desire correctly, and hence their failure to perform the act that they would otherwise have performed, isn't their fault. It isn't their fault because they lacked one of the capacities required for them to access or believe or instrumentally desire correctly, and hence to act correctly. They have an excuse. They could still have done so, of course. But if they had done so, then it would have been a complete fluke, and so not something for which they could take credit (compare Smith 2003).

But even though the standard story of action is economical, and promises to make relatively transparent the idea of there being rational requirements on actions, and the idea of our being responsible for what we do to the extent that it is our fault, it also promises to drive a wedge between actions that conform to rational requirements, on the one hand, and those that conform to moral requirements, on the other. Moral requirements would therefore still stand in need shoring up, as would specifically moral responsibility. Worse still, whatever we end up saying about moral requirements and moral responsibility, it now looms as a serious possibility that people may be morally required to act in certain ways even though we could rationally criticize them for acting in those ways, and that they could be also be morally responsible for acting in certain ways even though, if they were to act in those ways, they would have to be flawed from the rational point of view, either by being at fault or by being incapacitated in some way.

Why does the standard story drive this wedge? Consider again the questions we ask when we examine the rational status of an action. Are the beliefs that produce agents' actions well-supported by the evidence available to them? And are the strengths of their various instrumental desires proportional to the strengths of their background final desires and the probabilities that they assign to the objects of their instrumental desires being ways of bringing about what they finally desire? The problem is that, if Hume is right, questions like these exhaust those that bear on the rational status of actions, as there is no way to rationally criticize the *final* desires that produce actions. But if there is no way to rationally criticize the final desires that produce actions, then we cannot rationally criticize agents whose only "fault" is that they act on the basis of such final desires (compare Foot 1972).

Here is a passage in which Hume makes the crucial point about final desires.

> A passion is an original existence, or, if you will, modification of existence, and contains not any representative quality, which renders it a copy of any other existence or modification. When I am angry, I am actually possest with the passion, and in that emotion have no more a reference to any other object, than when I am thirsty, or sick, or more than five foot high. 'Tis impossible, therefore, that this passion can be oppos'd by, or be contradictory to truth and reason; since this contradiction consists in the disagreement of ideas, consider'd as copies, with those objects, which they represent (Hume 1740: 415).

What Hume says in this passage, translated into modern jargon, is that since final desires (in his terms, "original existences") do not purport to represent things to be the way they are, they aren't the sort of psychological state which can be true or false, and hence not the sort of state for which there can be reasons. Though he doesn't explicitly say why this is so in this passage, it is clear what the explanation is supposed to be.

Hume thinks that finally desiring something is just a matter of liking it for its own sake, and/or being disposed to bring it about in virtue of what it itself is like. The contents of these psychological states are therefore not propositions that purport to represent how things are; they are propositions that represent how the world would be if it were the way the agent would like it to be for its own sake, or how the world would be if it were the way the agent is disposed to make it in virtue of what it is in itself. But if final desires do not have contents that purport to represent things to be the way they are then they cannot be true or false, and if they cannot be true or false then they cannot be the sort of psychological state for which reasons can be given, because all there is to a consideration's being a reason, according to Hume, is its being a consideration that counts in favour of the truth of the proposition for which it is a reason, and hence in favour of the truth of the psychological state that has that proposition as its content.

Hume didn't shy away from this conclusion, he positively reveled in it, as is clear from the following infamous passage.

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter... In short, a passion must be accompany'd with some false judgement, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgement (Hume 1740: 416).

Imagine someone who acts in the most morally heinous way imaginable. Perhaps he knowingly and intentionally tortures a baby just for fun. Has he done something that there was a reason not to do, or which we can rationally criticize him for doing? Hume's answer is: Not necessarily. If he knew exactly what he was doing, and if he just so happens to have a final desire to have fun that is in no way hedged about so as to rule out the possibility of his having fun when that comes at the cost of the suffering of a baby, and if this desire is so strong that it outweighs everything else he cares about —perhaps it is the only thing that he finally desires— then, in Hume's view, he may not be rationally criticizable. Indeed, it may be that he would be rationally criticizable if he failed to torture a baby just for fun, as that would suggest either a failure to believe according to the evidence available to him, or a failure of instrumental coherence.

Here we see quite vividly why morality's foundations seem so desperately in need of shoring up and why the shoring up seems to require us to respond directly to Hume's argument. Absent skepticism about both moral requirements and moral responsibility, it seems that moral requirements must, in some way, reduce to rational requirements. But what exactly is the response to Hume's argument to be? Hume says: "In short, a passion must be accompany'd with some false judgement, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgement." In other words, though instrumental desires have *parts* for which there may be reasons, as there may well be reasons for or against having the beliefs that partially constitute them, the *wholes* of which these beliefs are parts —the instrumental desires themselves— cannot be had for a reason, or contrary to a reason, because the final desire component is not the sort of psychological state that can be had for a reason.

Some philosophers think that this is where Hume's argument goes wrong. Here, for example, is Derek Parfit.

> According to Objectivists, we have instrumental reasons to want something to happen, or to act in some way, when this event or act would have effects that we have some reason to want. As that claim implies, every instrumental reason gets its normative force from some other reason. This other reason may itself be instrumental, getting its force from some third reason. But at the beginning of any such chain of reasons, there must be some fact that gives us a reason to want some possible event as an end, or for its own sake. Such reasons are provided by the intrinsic features that would make this possible event in some way good (Parfit 2011: 91).

Parfit here asserts that there are reasons to have instrumental desires, and he derives from this the conclusion that there must be reasons to have final desires. His own view, for example, is that the intrinsic nature of well-being provides everyone with a reason to desire that there be more well-being rather than less for its own sake, without regard to whose well-being it is (Parfit 2011: 40). The person described earlier who finally desires to have fun, but where the fun is in no way hedged about so as to rule out the possibility of his having fun even if that comes at the cost of the suffering of a baby, thus most certainly fails to have a final desire that Parfit thinks there is reason to have. For he fails to finally desire that there be more well-being rather than less without regard to whose well-being it is.

There are two ways to understand what Parfit says about reasons to have final desires. According to the first —which, for the record, is not what Parfit has in mind— a consideration counts in favour of finally desiring some state of affairs in virtue of its being a consideration that counts in favour of the truth of the proposition that that state of affairs is finally good. This way of replying to Hume's argument grants that every reason counts in favour of the truth of some proposition, and hence is a reason for believing, but takes issue with the inference from this premise to the conclusion that there are no reasons for final desires. On this way of understanding what Parfit says, reasons for finally desiring something *inherit* their status as reasons from their being reasons that support the truth of the claim that that thing is finally good. What's needed, on this way of understanding what Parfit says, is thus some account of the relationship between facts about final goodness and final desires that would explain why reasons for the latter should inherit their status as reasons in this way from reasons for the former. Let's call this the 'Inheritance Thesis'.

According to the second way of understanding what Parfit says in this passage —and, for the record, this is what he has in mind— we are committed to supposing that the intrinsic features of the things we finally desire provide reasons for finally desiring those things, but the concept of a reason in play here, and the concept of a reason for believing too for that matter, is a *primitive* concept. It cannot be further explained by anything. In particular, it cannot be explained in terms of truth in the way that Hume proposes. The upshot is that when Parfit says that "[s]uch reasons are provided by the intrinsic features that would make this possible event in some way good", this is not supposed to be in any way explanatory, but merely notes the connection between reasons for final desires and final goodness. On this way of understanding what Parfit says, Hume's argument fails because it contains a false premise about the nature of reasons, the premise that the concept of a reason can be explained in terms of the concept of truth.

The problem with this second way of understanding what Parfit says, however, is that it beggars belief to think that we cannot explain what it is for a consideration to be a reason for belief. Hume himself thought something much more specific than that we could explain what such reasons are in terms of the concept of truth. He thought that we could explain what reasons are in terms of the concept of *entailment*. A fact is a reason for believing, he thought, just in case that fact entails the truth of the proposition believed. Hume therefore happily embraced the radical conclusion that all reasons for belief are deductive reasons, and hence that there are no inductive or abductive reasons for believing. Many have baulked at this conclusion, of course, and quite rightly so. But reductionism about reasons for belief doesn't require us to buy into this radical Humean conclusion, not even a kind of reductionism that takes its inspiration from Hume's idea that the concept of a reason reduces to the concept of entailment.

The best way to develop a more modest Humean view in a way that still allows for the possibility of inductive and abductive reasons would to build on David Lewis's contextualist theory of knowledge (Lewis 1996). We might suppose that some fact p is a reason for a subject to believe that q if and only if and because p is the sort of thing that *could* give a subject knowledge that q, where this in turn is explained by the fact that, in those possible worlds in which the subject does know that q on the basis of p, the fact that p removes all of the other possibilities except for q that the subject isn't properly ignoring, where the norms of proper ignoring are semantic norms telling us when someone's forming a belief in the ignorance of certain facts counts as knowledge. Inductive and abductive reasons would both be possible, according to this view, though their status as reasons would be contingent on what's properly ignored. Would a view like this explain the concept of a reason? It most certainly would, as it would spell out quite precisely what relationship a consideration has to stand in to the content of a belief if that consideration is to count as a reason. The idea that the concept of entailment gives us no purchase on what a reason is should therefore be rejected out of hand, and with it the idea that the concept of a reason is a primitive concept.

Reductionism about reasons for belief of this kind is, however, very bad news for Parfit. For on the assumption that we aren't equivocating when we talk of reasons for believing and reasons for desiring, it follows that what he says in the passage makes sense only if the concept of a reason for final desiring similarly reduces to the concept of entailment. But Hume's argument purports to show that it does not, as a final desire doesn't purport to represent things to be the way they are. So if Hume's argument stands —and at this stage we should think that Parfit has given us no reason to suppose that it doesn't— then the only conclusion to draw would be that, though there are reasons for the belief component of our instrumental desires, there are no reasons for the final desire component, and hence no reasons for instrumental desires.

Of course, even if Hume's argument does stand, agents are still required to be instrumentally coherent, so we can still rationally evaluate instrumental desires independently of whether their belief component or final desire component are had for reasons. But since instrumental coherence takes agents' final desires as given, and merely constrains how strong agents' instrumental desires are to be by the strengths of their final desires and the likelihood that they attach to the objects of their instrumental desires making the world they finally desire it to be, instrumental coherence allows us to rationally evaluate instrumental desires without supposing that there are reasons for desires, whether final or instrumental.

This leaves us with the first way of understanding what Parfit says in the passage quoted. The first idea, you will recall, is that reasons for finally desiring are such reasons *in virtue of* being reasons that support the truth of the claim that the object of the final desire is finally good. On this way of understanding what Parfit says, Hume is right that all reasons are reasons for belief, but wrong that it follows from this that there are only reasons for belief. This is the Inheritance Thesis. However, as I said, this way of understanding what Parfit says clearly isn't what he has in mind. The reason I said that is because Parfit thinks, with Thomas Scanlon (1998), that facts about final goodness are facts about reasons for having certain final desires.

> When we call something *good,* in what we can call the *reason-implying* sense, we mean roughly that there are certain kinds of fact about this thing's nature, or properties, that would in certain situations give us or others strong reasons to respond to this thing in some positive way, such as wanting, choosing, using, producing, or preserving this thing... Things can be good or bad in other senses... But the most important uses of 'good' and 'bad' are, I believe, reason-implying (Parfit 2011: 38).

To believe that something is finally good, according to Parfit, is therefore just to believe, *inter alia*, that it has certain intrinsic features that provide us with a reason to finally desire it (Parfit 2011: 50). But if this is right then reasons for finally desiring something cannot be reasons *in virtue of* being reasons that support the truth of the claim that the thing is finally good, as this would get the order of explanation the wrong way around. Parfit must therefore reject the Inheritance Thesis.

It doesn't follow that the Inheritance Thesis *isn't* true, of course. All that follows is that, if it is true, we cannot accept the Scanlon-Parfit view that goodness is just a matter of what there is reason to desire. Is there any independent reason to think that the Inheritance Thesis is true? What view of goodness does it presuppose? And, if that view of goodness is plausible, and we do come to accept the Inheritance Thesis as a result, does this provide us with the wherewithal to resist Hume's views about the rational status of final desires? To answer these questions, we need to start much further back. We need to remember some lessons taught to us by Judith Jarvis Thomson in her wonderful, if flawed, *Normativity* (Thomson 2008, Smith 2010).

Thomson points out that many kinds of things are what she calls 'goodness-fixing kinds'. These are kinds whose nature fixes a standard of success for things of that kind. Here are some examples. *Toasters* are devices for warming and browning bread so that you can enjoy eating it, so a good toaster is device that does all of this without burning the toast, so making it much more enjoyable to eat. *Burglars* are people who make their living by breaking into buildings and stealing things, so a good burglar is someone who reliably does this without getting caught. *Tennis players* are people who play tennis, so a good tennis player is someone who reliably wins all of his games. Goodness-fixing kinds contrast with those kinds that don't fix a standard of success. Thomson's examples of these are *pebble, smudge,* and *cloud* (Thomson 2008: 21-22). No two pebbles, or smudges, or clouds can differ in that one is better at being a pebble or a smudge or a cloud than the other, in virtue of the standards that are internal to these kinds. Not so for two toasters, burglars, or tennis players.

Once we notice that there are such kinds, it should immediately strike us that the kind *agent* is a goodness-fixing kind. Think again about the standard story of action. Someone is an agent in virtue being capable of action, which is to say, in virtue of having the capacity to realize their final desires, given their beliefs. This fixes a standard of performance for agents. A good agent is one who has and exercises, to a high degree, the capacity to form beliefs about the world in which he lives in the light of reasons and to realize his final desires in it. Since an agent's beliefs about how to realize his final desires have to be not just had for reasons, but also true, if he is to realize them in action, we can restate what it is for someone to be a good agent. A good agent is someone who has and exercises, to a high degree, the capacity to know the world in which he lives and to realize his final desires in it.

The Dispositional Theory of Value in effect uses the fact that *agent* is a goodness-fixing kind in order to provide an analysis of a different concept of final goodness (Smith 1994, Smith 2010). According to Dispositional Theory, what it is for something to be finally good in this different sense, as indexed to some agent A, is for that thing to be the object of a final desire that A's *maximally good* counterpart has. There are thus two quite distinct concepts of goodness in play. The latter concept of goodness is the one internal to goodness-fixing kinds. The former is the one that we have defined in terms of the latter.

To avoid confusion, let's use the term 'ideal' to pick out a maximally good member of a goodness-fixing kind and let's use 'good' to name the other property of goodness. In these terms, the Dispositional Theory tells us that an outcome is finally good, as indexed to an agent A, just in case A's idealized counterpart —that is, A himself in the nearest possible world in which he has and exercises a maximal capacity to know the world in which he lives and realize his intrinsic desires in it— finally desires that outcome. For A to believe an outcome to be finally good, as indexed to him, is thus just for him to believe that his idealized counterpart finally desires it.

The attractions of the Dispositional Theory in explaining the Inheritance Thesis is, I hope, clear. According to the Inheritance Thesis, reasons for finally desiring something inherit their status as reasons from their being reasons that support the truth of the proposition that that thing is finally good. As we saw earlier, what we need in order to make this idea seem plausible is some account of the relationship between final goodness and final desire, an account that explains why reasons for final desires should inherit their status as reasons in this way. The Dispositional Theory provides the needed account because it tells us that facts about final goodness, as indexed to an agent, are fixed by the facts about the final desires of that agent's ideal counterpart. These are the final desires that an agent should have, in the sense that his having those final desires is required for him to meet the highest standards that are internal to the concept of agency. The contents of an ideal agent's final desires and the objects her knowledge about what's finally good are therefore identical.

The Dispositional Theory, together with the Inheritance Thesis, thus suggest that beliefs and desires are far more similar to each other than Hume would have us think (Pettit and Smith 1996). Just as there is a sense in which an agent's beliefs should have as their contents the contents of the knowledge of that agent's idealized counterpart, so there is a sense in which an agent's desires should have as their contents the contents of the desires of the agent's idealized counterpart. The former is the kernel of truth in the claim that belief aims at the truth. The latter is the kernel of truth in the claim and desire aims at the good. Despite this similarity, however, and despite the fact that we have now found decisive reasons to reject Hume's claim that there are no reasons for desires, note that we have still found no reason whatsoever to reject his account of the rational evaluation of final desires.

According to the Dispositional Theory, the final desires that would be possessed by an agent's ideal counterpart are fixed by the norms internal to the concept of agency. However nothing we have said so far gives us any reason to suppose that these norms go beyond those that Hume proffers. But if Hume's account of the norms governing beliefs and desires is right, then an agent's ideal counterpart is going to possess whatever final desires the non-ideal agent possesses. What will be finally good, relative to each agent, will simply be that the realization of that agent's final desires. To return to the example that worried us earlier on, for all that's been said, it may therefore be finally good, relative to some agent, that he has fun even when that comes at the cost of torturing babies. Though establishing the truth of the Inheritance Thesis is necessary for us to take issue with Hume's account of the norms governing final desires, it plainly isn't sufficient for us to do so.

As I understand it, it is at this point in the dialectic that a Constitutivist Theory of Reasons comes to the fore. Constitutivists buy into everything that has been said thus far, but they add one crucial qualification. They insist that Hume's characterization of an ideal agent is inadequate because he fails to see that certain final desires are *constitutive* of what it is to be an ideal agent. More precisely, they think that *all* ideal agents have certain dominant final desires in common, where these desires are dominant in the sense that their realization is a condition of the realization of any other desires that an ideal agent might happen to have. The final desires that are constitutive of being ideal therefore make it the case that certain things are finally good no matter which agent final goodness is indexed to.

The Constitutivisits's account of the dominant final desires that are constitutive of being an ideal agent thus provide the much needed link between rational requirements and moral requirements that we've been looking for. In conjunction with the Inheritance Thesis and the standard story of action, it entails that there are certain final desires that everyone has reason to have, and so certain actions that everyone has reason to perform, and it further entails that agents with the requisite rational capacities are responsible for failing to have these dominant final desires and performing these actions when their failure to do so is a result of their failure to exercise these capacities, and it identifies these actions with those that are morally required. All this and much more is thus the promise of a Constitutivist Theory of Reasons. The question that remains is how Constitutivists manage to deliver on this promise.

## 2. PARTS

Though a Constitutivist Theory of Reasons can take many different forms, in what follows I will focus on what seems to me to be the most promising version of the theory. For brevity, I will call this view 'Constitutivism' and I will call the theorist who advances it 'the Constitutivist'. However, it is important to remember that other theorists do defend different versions of the theory, so problems that might arise for the version proposed here may not be problems for them (see again Korsgaard 1996, Velleman 2005). Having said that, let me reiterate that the version I go on to describe does seem to me to be the most plausible and powerful version, and that my firm hunch is that it avoids the problems that face other versions. That is why I say that the version described here seems to me to be the most promising.

Having said that, let me emphasize that my main aim in what follows is not to mount a full defence of Constitutivism, but rather to describe the main parts of the theory in the hope that doing so will make it clear what its virtues are (for a partial defence see Smith 2011, Smith 2012). The division of the theory into parts will provide a framework for comparing different versions. The parts are: (i) a diagnosis of the main problem facing Hume's account of an ideal agent; (ii) an explanation of how that diagnosis leads to the conclusion that certain desires are constitutive of being an ideal agent; (iii) an explanation of how that provides us with an account of what is morally required at the most fundamental level, and how this account dovetails with an account of moral responsibility; and (iv) a derivation of various subsidiary moral principles and subsidiary responsibilities.

### 2.1. Diagnosis of the main problem facing Hume's account of an ideal agent

The Constitutivist's first and most important task is to diagnose where Hume's account of an ideal agent goes wrong. According to Hume, an ideal agent is one who fully and robustly possesses and exercises the capacities to do two things: to have knowledge of the world in which he lives, and to realize his desires in it. The main problem with this account, according to the Constitutivist, is that in a wide range of circumstances their exercise pulls in opposite directions. The full and robust exercise of the one capacity does not fully cohere with the full and robust exercise of the other. To the extent that this is so, the ideal agent's psychology is therefore not *maximally* coherent. Since it is a contradiction in terms to suppose that an ideal agent's psychology is not maximally coherent —two psychologies that differ in that one is more coherent than the other are also such that the more coherent one is more ideal— Hume's account of an ideal psychology must therefore be mistaken.

In order to see why all of this is so, two important points must be kept in mind. The first point is that an agent who fully and robustly exercises the capacity to have knowledge of the world in which she lives is one who exercises that capacity across a wide range of possible circumstances, including those in which she has very different final desires from those she actually has, and an agent who fully and robustly exercises the capacity to realize her desires is similarly one who exercises that capacity across a wide range of possible circumstances, including those in which she knows very different things about the world from those things she actually knows, given that the facts about the world differ from world to world. The second is that, according to Hume's account of desire, an ideal agent can have final desires for anything. This is because being ideal does not in any way constrain the contents of an ideal agent's final desires.

With these two points in mind, consider the bare possibility that an agent finally desires that she now believes that p. This agent's ideal counterpart desires that she now believes that p and, as she fully and robustly exercises the capacity to realize this desire, this in turn means that she must believe that p whether or not p is true. This is what we learn from the second point. But since in order to be ideal, she also has to fully and robustly exercise the capacity to know the world in which she lives, it follows that in the formation of her beliefs about p, she must also be sensitive to whether or not p is true. This is what we learn from the first point. So every agent's ideal counterpart is both sensitive to whether or not p is true in forming the belief that p, and believes that p whether or not p is true. Since this is a contradiction in our description of an ideal agent, a contradiction to which we have found ourselves committed by working through the implications of Hume's account of an ideal agent, the account must be rejected.

The Humean must say that we have somehow misdescribed their account of an ideal agent. The Humean could try suggesting that an ideal agent doesn't have to fully and robustly possess and exercise the capacity to realize his desires, only the capacity to know the world in which he lives, but in what sense would this make him be an ideal *agent*? Or the Humean could try suggesting that an ideal agent doesn't have to fully and robustly possess and exercise the capacity to know the world in which he lives, only the capacity to realize his desires in it, but then how would he be imagining that the ideal agent manages to realize his desires? Or —and this is the most plausible response for the Humean to give— he could try suggesting that an ideal agent does have to fully and robustly possess and exercise both of these capacities, but insist that since the Constitutivist is right that their exercise pulls in opposite directions from each other, a maximally coherent agent is one who is assessed as being such separately along two quite different dimensions: knowledge acquisition and desire-realization. An ideal agent thus turns out to be one whose psychology, by its very nature, displays lots of tension and disunity, as a higher score along one dimension comes as the cost of a lower score along the other.

But though this is the most plausible response for the Humean to give to the problem identified with his account, we should go along with what he says only if no alternative account of what it is to be an ideal agent reduces the amount of tension and disunity inherent in the Humean's conception of an ideal agent's psychology. In particular, we should go along with it only if there are no additional psychological states that are plausibly thought of as being constitutive of an ideal psychology, much an agent's possession and exercise of the dual capacities to know the world in which he lives and realize his desires in it are plausibly thought to be constitutive of an ideal psychology, but possession of which would ensure that an ideal agent's psychology is much more coherent and unified than it is according to the Humean's conception. If there are such psychological states, then we should suppose instead that they too are partially constitutive of an agent's being ideal.

### 2.2. Explanation of how the diagnosis leads to the conclusion that certain desires are constitutive of being an ideal agent

The Constitutivist thinks that there are such additional psychological states. In the most abstract terms possible, these are those psychological states, whatever they are, possession of which, alongside the dual capacities to know the world in which an agent lives and realize his desires in it, induce more coherence and unity in his psychology. Less abstractly, the Constitutivist's suggestion is that these psychological states are *coherence-inducing desires*. The argument he gives for this conclusion is an argument to the best explanation. Certain coherence-inducing desires are such that their possession would induce more coherence and unity in an agent's psychology; the Constitutivist can think of no other psychological states that could play this role equally well; so he concludes that these coherence-inducing desires do play this role.

Imagine that all ideal agents have a dominant final desire that they do not now interfere with their current exercise of their capacity to have knowledge of the world, where a dominant desire is one that overrides all of the desires that aren't partially constitutive of an agent's being ideal, including desires like the desire to believe that p. This would remove all potential for conflict between the full and robust exercise of an agent's capacity to know the world in which she lives and the full and robust exercise of her capacity to realize her desires —call these 'potentially idiosyncratic' desires. The only way in which an agent who desires to believe that p could robustly possess and exercise the capacity to realize her desires in worlds in which she is otherwise ideal, would be by leaving herself free to exercise the former, as her desire not to interfere would dominate her desire to believe that p. Since an ideal agent would have and exercise the capacity to know the world in which she lives, this means that she would wind up knowing that p, or knowing that not p, depending on whatever happens to be the case in the world in which she lives.

Once we see that an agent's final desire to not now interfere with her exercise of her capacity to have knowledge of the world in which she lives is plausibly thought to be constitutive of being ideal, the Constitutivist thinks that other desires can be seen to be similarly constitutive of her being ideal. For example, consider someone who is otherwise ideal at a time, but not such as to be ideal at later times, and compare him to someone who is otherwise ideal at a time, and also such as to be ideal at later times. Which of these agents is more ideal at a time? For example, consider someone who wants to believe that p now, and someone else who wants to believe that p later. Both are equally such that, in order to exercise their capacity to realize their desires now, they have to interfere with their exercise of their capacity to have knowledge of the world, the one now, the other later. Is one of these agents more ideal than the other?

The Constitutivist thinks that these agents are each equally less than ideal, and that what this shows is that being ideal now requires an agent to be such as she needs to be in order to be ideal not just now, but also later. Of course, an agent's being ideal later depends on something that the agent cannot control now, namely, whether or not she later exercises her capacities. But this leaves a great deal that she can control now through the direct exercise of her capacities. She can leave herself free to exercise her capacities later, and she can make sure that she has capacities later to exercise. The upshot is that, in addition to a dominant desire that they do not now interfere with their current exercise of their capacity to have knowledge of the world, ideal agents must have two further dominant desires as well: a dominant desire that they do not now interfere with their later exercise of their capacity to have knowledge of the world, and a dominant desire they do what they can now to ensure that they have the capacity to have knowledge of the world later to exercise.

So far we have focused on desires that ensure that an ideal agent can exercise her capacity to have knowledge of the world. But everything that's been said so far about the threat that an agent's potentially idiosyncratic desires present to her possession and exercise of her capacity to have knowledge of the world, whether now or later, applies equally to the threat that potentially idiosyncratic desires present to an agent's possession and exercise of her capacity to realize her desires (on condition of course, that the realization of those desires doesn't require her interfering with the possession or exercise of her capacities —I will take this qualification as read in what follows).

Imagine, for example, an agent who now finally desires that p will be the case later, whether or not she will later desire that p is the case, and so lays traps for her future self to ensure the frustration of any inconsistent desires, should she acquire them. Or imagine an agent who is now totally indifferent to the fact that she will later lack the capacity to realize her desires, or perhaps one who positively desires that she lacks that capacity. These agents are also less than ideal in virtue of not being now such as to be ideal later. The upshot of these parallels, according to the Constitutivist, is that ideal agents must also have dominant final desires that they do not interfere with their current or later exercise of their capacity to realize their desires, and to do what they can to ensure that they have the capacity to realize their desires later.

Finally, the Constitutivist thinks that the limitation of these final desires contents to an agent's own present and future possession and exercise of her capacities is *ad hoc*. An ideal agent would desire not to interfere with the exercise of the knowledge-acquisition and desire-realization capacities of not just herself in the present and the future, but of anyone whose possession and exercise of their knowledge-acquisition and desire-realization capacities is dependent on what she currently does. A variety of arguments can be given for this conclusion, but the most powerful is a symmetry argument (compare Parfit 1984, especially §55).

Though, as we have seen, there is a deep difference between an agent's relationship to her own current beliefs and desires, and those she has later, given that her current beliefs and desires are the ones she directly controls through the exercise of her rational capacities, there is no such deep difference between her relationship to her own later beliefs and desires and those of other people. The upshot is therefore that, just as an agent's being ideal at a time requires her to be, at that time, such as she needs to be in order to be ideal, not just at that time, but also at later times, so her being ideal at a time requires her to be, at that time, such as she needs to be in order to be ideal not just herself, at that time and at later times, but also as she needs to be for others to be ideal, whether at that time or at later times.

Let's sum up. Constitutivists argue that each ideal agent has the following final desires in common: the desire not to interfere with the current or future knowledge-acquisition or desire-realization capacities of any being whose exercise of these capacities is dependent on what the agent herself currently does, and the desire to do what she can to ensure that those whose possession of such capacities is dependent on what she currently does have such capacities to exercise. For short, let's call these the desires to help and not interfere. Constitutivists further insist that the realization of these desires is a condition of the realization of all other potentially idiosyncratic desires that an ideal agent might happen to have, and hence that they must be dominant. They must be dominant because only so could they play their crucial coherence-inducing role.

## 2.3. Explanation of how the fact that certain desires are constitutive of being ideal provides us with accounts of both the most fundamental moral requirements and the conditions of moral responsibility

In order to turn the Constitutivist's account of the desires that are constitutive of being an ideal agent into an account of moral requirements, we need to add to that account the materials adduced earlier: the Dispositional Theory of Value, the Inheritance Thesis, and the account of the rational standing of action suggested by the standard story of action. Once we have an account of moral requirements in hand, we can provide an account of the conditions of moral responsibility.

Let's begin with the account of moral requirements. Given that final goodness, as indexed to an agent, is fixed by what that agent's ideal counterpart desires —this is what we learn from the Dispositional Theory of Value— and given that helping and not interfering are finally desired by every agent's ideal counterpart —this is what Constitutivism tells us— it follows that helping and not interfering are finally good no matter to which agent final goodness is indexed. When we combine this conclusion with the Inheritance Thesis —this is the claim that reasons for finally desiring something inherit their status as reasons from their being considerations that support the truth of the proposition that that thing is finally good— the upshot is that everything that's been said so far constitutes a reason for agents to finally desire to help and not interfere. And when we combine this conclusion with the standard story of action's account of the rational standing of actions in terms of the rational standing of the beliefs and desires that produce those actions, the upshot is that every agent has a reason to help and not interfere.

The significance of this last conclusion cannot be overstated. We saw earlier that the rational evaluation of agents' actions requires us to ask questions like these: Are the beliefs that produce agents' actions well-supported by the evidence available to them? Are the strengths of their various instrumental desires proportional to the strength of their background final desires and the probabilities that they assign to the objects of their instrumental desires being ways of bringing about what they finally desire? But what we've learned from Constitutivism, the Dispositional Theory of Value, and the Inheritance Thesis, is that we must ask another question as well. We must ask whether the dominant final desires that move agents to act are those that they have reason to have. More specficially, since agents all have reasons to have the same desires, we must ask whether they are moved to act by the desires to help and not interfere.

If these questions get positive answers, then the beliefs and final desires that produce agents' actions, and the actions they produce, conform to rational requirements. In every case, whatever else agents are doing, it turns out that these will be acts of helping and not interfering. The striking similarity of these acts to those that we ordinarily take to be morally required is, the Constitutivist insists, manifest. The only reasonable conclusion to draw is thus that every agent isn't just rationally required to help and not interfere, but that, at the most fundamental level, every agent is morally required to help and not interfere as well. The concern that we may be morally required to act in ways we have no reason to act is thereby laid to rest.

Note that this doesn't just provide morality with the rock solid foundation in rational requirements that it so sorely needs, but that it also provides us with a plausible and intuitive account of the conditions under which agents can held morally responsible for their actions. Imagine some agent who fails to help and not interfere. Is he responsible for having acted wrongly? The Constitutivist's answer is that he is responsible to the extent that he had, but failed to exercise, the capacity to recognize and respond to arguments like those provided here. For any agent who fails to help and not interfere while having that capacity is someone whose failure to acquire the desires to help and interfere, and so his failure to act on these desires, is traceable to his failure to exercise his capacity to recognize and respond to relevant arguments. There-in lies his fault. Incapacity would of course excuse, and difficulty in the exercise of an agent's capacities may mitigate his fault, but absent excuse or mitigation, those who act wrongly have no one to blame for their wrongdoing but themselves.

### 2.4. Derivation of subsidiary moral principles and subsidiary responsibilities

So far we have focused on what is morally required at the most fundamental level. Importantly, however, note that the account provided suggests that there will be many subsidiary moral principles. Let me briefly describe what some of these subsidiary moral principles might be, just to convey some sense of how powerful the Constitutivist's account of the most fundamental moral requirements really is.

Imagine a situation in which one agent makes a promise to another, but then knowingly fails to keep that promise without taking steps to warn her, and without having some compelling reason to do something else instead. For example, suppose a young man promises to meet his friend at the movies, and she turns up at the agreed time, but he fails to show up, not because he was (say) tending to the victim of a traffic accident that he had on the way to the movie, but because he just didn't feel like going when the time came. What is the wrong that he does, exactly?

The Constitutivist's account of the most fundamental moral requirements suggests an answer. The young man knowingly interferes with his friend's exercise of her capacity to have knowledge of the world in which she lives, as he led her quite reasonably to believe, falsely, that he would be at the movie, and he also knowingly interferes with her exercise of her capacity to realize her desires, as he led her to form that belief knowing that she would act on her desire to spend the evening with him, a desire that she had no chance of realizing, given that he wouldn't be there. He therefore acted in the knowledge that, if she hadn't falsely believed that he would be there, she would have acted on some other desire instead, a desire that she would have at least had a chance of realizing.

The requirement to keep promises is thus a clear example of a subsidiary moral requirements, according to Constitutivism, and with this example in clear view, we can see that there will be a whole host of other subsidiary moral requirements as well, subsidiary moral requirements grounded in the fact that the creation of reasonable expectations is ubiquitous in human life. Wherever such reasonable expectations are created, but go unmet, the account provided thus suggests that there is a wrong in the offing of the most fundamental kind. Lying, manipulating, cheating, being disloyal, betraying, and free-riding are all examples of subsidiary wrongs of this nature. Moreover, wherever there are subsidiary wrongs of these kinds, and the agents of those subsidiary wrongs had the capacity to recognize and respond to the arguments for acts of these kinds being subsidiary wrongs, the agents of such acts will be at fault.

Constitutivism suggests that there will be another quite different class of subsidiary moral requirements as well. Imagine a father who fails to provide his child with the necessary lessons in life to be able to access and rationally evaluate the evidence available to her as she forms her beliefs about the world in which she lives, and who also fails to equip her with the personal resilience and confidence required to create opportunities for herself, as she goes through life. Perhaps he drums into her his own views that books are full of useless information, that no one outside the family is to be trusted, that the only way for a girl to get ahead is by being attached to some powerful man, that she is unattractive and that no man will ever be attracted to her, and so on. What exactly is the wrong that a father like this does to his child?

The Constitutivist's account of moral requirements once again suggests the basics of an answer. The child is, after all, someone whose development of her capacities to have knowledge of the world in which she lives and realize her desires in it is dependent on what her father does for her as he raises her, and he not only fails to do all that he can to ensure that she has such capacities to realize, he ensures that the capacities she has are warped in all sorts of ways. He was subject to a subsidiary moral requirement to teach her that books are full of useful information, that many people outside the family are to be trusted, that there are many ways for girls to get ahead without being attached to some powerful man, and so on, but these are all subsidiary requirements that he violates. Assuming that he had the capacity to recognize and respond to the arguments that might be given for these being subsidiary moral requirements, he is therefore also at fault for his violations.

The situation of the father is hardly unique. We are all in a position to influence the development of others' capacities to know the world in which they live and realize their desires in it, so the Constitutivist's account of what's morally required at the most fundamental level suggests that there will be a whole host of such subsidiary moral requirements. Some of these will be as banal as helping those who are clearly lost to find their way, encouraging them to put their trust in others without thereby losing their faith in themselves, and so on. Others will be much more significant, such as taking steps to ensure that everyone has access to a proper education and the opportunity to live in circumstances of political and social equality. To the extent that each of us has the capacity to recognize and respond to the arguments that can be given for these being subsidiary moral requirements, we are therefore all at fault to the extent that we fail to exercise that capacity. So, at any rate, Constitutivism suggests.

## 3. CONCLUSION

We saw at the outset that the promise of a Constitutivist Theory of Reasons is that it can help us do something that we desperately need to do, which is to reduce facts about the wrongness of actions to facts about the rational requirements to which actions are subject. Constitutivism does this by dividing that task into distinct parts. It begins by offering a compelling diagnosis of the problem facing the Humean, anti-rationalist, account of an ideal agent. It shows how that diagnosis leads to the conclusion that the desires to help and not interfere are constitutive of being an ideal agent. It explains how the fact that that the desires to help and not interfere are constitutive of being an ideal agent provides us with all we need to give an account of what is morally required at the most fundamental level, and how this account in turn dovetails with a plausible account of moral responsibility. And, finally, it derives various subsidiary moral principles and subsidiary responsibilities from its account of what's morally required at the most fundamental level. Much work still needs to be done in filling out the details, of course. But hopefully enough has been said to make it clear how well-placed Constitutivism is to deliver on its promise.

## BIBLIOGRAPHY

Blackburn, S., 1998: *Ruling Passions*, Oxford: Clarendon Press.

Davidson, D., 1963: "Actions, Reasons, and Causes", reprinted in his *Essays on Actions and Events,* Oxford: Oxford University Press, 1980.

— 1971: "Agency", reprinted in his *Essays on Actions and Events,* Oxford: Oxford University Press, 1980.

Enoch, D., 2006: "Agency, Shmagency", *Philosophical Review* 115: 169-198.

— 2011: "Shmagency Revisited", in *New Waves in Metaethics*, ed. Brady, M., London: Palgrave Macmillan.

Foot, P., 1972: "Morality as a System of Hypothetical Imperatives", reprinted in her *Virtues and Vices*, Berkeley: University of California Press, 1978.

Hume, D., 1968 [1740]: *A Treatise of Human Nature,* Oxford: Clarendon Press.

Kant, I., 1948 [1786]: *Groundwork of the Metaphysics of Morals,* London: Hutchinson and Company.

Korsgaard, C., 1996: *The Sources of Normativity*, Cambridge, UK: Cambridge University Press.

Lewis, D., 1996: "Elusive Knowledge", *Australasian Journal of Philosophy* 74: 549-567.

Nagel, T., 1970: *The Possibility of Altruism*, Princeton: Princeton University Press.

Parfit, Derek, 2011: *On What Matters: Volume One.* Oxford: Oxford University Press.

Pettit, P. and M. Smith, 1996: "Freedom in Belief and Desire", *Journal of Philosophy* 93: 429-449

Scanlon, T., 1998: *What We Owe To Each Other*, Harvard: Harvard University Press.

Smith, M., 1994: *The Moral Problem,* Oxford: Wiley-Blackwell.

— 2003: "Rational Capacities", in *Weakness of Will and Varieties of Practical Irrationality*, ed. Stroud, S. and C. Tappolet, 17-38, Oxford: Oxford University Press.

— 2004: "The Structure of Orthonomy", in *Action and Agency* (Royal Institute of Philosophy Supplement: 55) ed. Hyman J. and H. Steward, 165-193, Cambridge: Cambridge University Press.

— 2010: "On *Normativity*", *Analysis Reviews* 70: 715ñ731.

— 2011: "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values", *Ratio* 24: 351-363.

— 2012: "Four Objections to the Standard Story of Action (and Four Replies)", *Philosophical Issues: Action Theory* 22: 387-401.

— Forthcoming: "Agents and Patients, Or: What We Learn about Reasons for Action by Reflecting on Process-of-Thought Cases" in *Proceedings of the Aristotelian Society*.

Velleman, D., 2005: *Self to Self,* New York: Cambridge University Press.