Explaining Reasons

Stephen Finlay

This is a preprint of an article to be published in the *Deutsche Jahrbuch für Philosophie*, Vol. 3.
Felix Meiner Verlag, 2012.

What does it mean to call something a "reason"? Here I offer a unifying semantics for the English word 'reason' which purports to account for the range of different ways we ordinarily use it (and, I hope, extends to counterparts in other languages, like 'grund' in German). This account challenges three ideas that are popular in contemporary philosophy. The first is the idea that 'reason' is semantically ambiguous, having several distinct meanings. The second is the idea that our concept of a *normative* reason is the basic normative concept in terms of which all other normative concepts must be analyzed. I argue instead that we can analyze talk about normative reasons by appeal to a more basic normative concept of *goodness*. The third idea is that the basic normative concepts are primitive, and cannot be reductively analyzed into entirely nonnormative components. I show how a number of apparent obstacles for the proposed analysis can be overcome if we adopt a reductive, *end-relational* analysis of the meaning of 'good', which I have championed elsewhere.[1]

1. Reasons as Explanations Why

Philosophers often take talk about "reasons for action" as their object of study, which is to focus narrowly on a subset of the ordinary ways we use the word 'reason'. The word could just be semantically ambiguous. But before concluding this we should investigate the simpler unifying hypothesis that these are all different uses of the same word with a single meaning. I shall argue that this hypothesis can be vindicated, casting light on the nature of normative reasons. (However, I set aside 'reason' in its uses referring to mental faculty.)

---

[1] This paper is based on a draft chapter of my manuscript on normative language, *Confusion of Tongues*, and develops some ideas in Finlay 2001, 2006. Constraints of space and time force my references and acknowledgments here to be thin. Among many to whom I owe thanks are audiences at the XXII. Deutscher Kongress für Philosophie, UC-Davis, and Reed College. I am especially grateful to Thomas Spitzley and the DGPhil for the opportunity to present this paper in Munich.

A first expansion of the data comes from noticing that talk about "reasons for action" is only part of our talk about normative reasons. We also talk about reasons to *want* things, reasons to *hope for* things, reasons to *regret* things, reasons for anger, guilt, fear, etc. These are just a few examples of reasons *for attitudes*, among which we can also count the important case of reasons for *belief*. A unifying semantics for 'reason' must account for all these cases. (Some philosophers claim that reasons for action are properly analyzed also as reasons for attitudes: by 'a reason to φ' we really mean *a reason to intend to φ*. This theoretically motivated view finds no support in linguistic practice or the analysis offered here.)

A second expansion of the data arises from distinguishing between two different kinds of "reason for action". In addition to talking about ("objective") normative reasons, which are facts that favor actions or attitudes ('reasons to φ'), we talk about the reasons that motivate agents to φ, the 'reasons *for which s* φ's, and '*s*'s reasons for φ-ing'. Philosophers have drawn a sharp distinction between a normative and a motivating sense of 'reason for action', and relatedly, between an "objective" and a "subjective" sense. This presents a challenge for a unifying semantics.

Our data expands in a third, more dramatic way when we notice that 'reason' is commonly used to talk about (nonnormative) reasons for ordinary *facts*; for example

(1) 'The reason the light isn't turning green is that the car didn't cross the sensor.'
(2) 'The reason this escape plan won't work is that the prison fence is electrified.'

As this use of 'reason' is plausibly the most common and transparent, I start my search for a unified meaning here, and will then examine how far this sense of 'reason' can be extended.

In these cases, by "a reason" we refer to a fact *that r* (e.g. that the car didn't cross the sensor) that stands in a relation *R* to another fact *that p* (e.g. that the light isn't turning green). In looking for what we *mean* by 'reason' in these uses, the answer seems obvious: an *explanation why*. The relation *R* is just the relation of being explanatory of why, and for *r* to be a reason for *p* is simply for *r* to be the explanans to *p*'s explanandum. Or as we may say, a reason is an answer to a 'why?' question; 'Because…'. The sentences above seem to be equivalent in meaning to

(1') 'The explanation why the light isn't turning green is that the car didn't cross the sensor.'

(2') 'The explanation why this escape plan won't work is that the prison fence is electrified.'

While the concept of explanation itself raises philosophical puzzles, here I'll assume the view that 'explanation' means *something that makes clear* (or *reveals*), and 'why' means *what makes it true that*. Saying that *r* is the reason that *p* is therefore saying that *r* is something that reveals what makes it true that *p*. So (e.g.) the fact that the car didn't cross the sensor is something that reveals what makes it true that the light isn't turning green.

2. Reasons to φ

Can our talk about normative reasons to φ be analyzed in terms of explanations why? There is nothing novel about this idea, which has been endorsed by many philosophers. But it has proved difficult to implement, and in this section I propose a new strategy.

One challenge is to reconcile the grammar of normative reasons sentences with the logical form *R*(*r*, *p*). A simple report of a normative reason may take the form, 'That *r* is a reason to φ,' as in

(3) 'That plaque decays teeth is a reason to brush your teeth every day.'
(4) 'That the subsequent train isn't until 11:30am is a reason to catch the 8:30am train.'

Here, 'reason' grammatically takes a 'to φ' complement, prompting some to claim that normative reasons involve relations to actions rather than to propositions. We can of course ask, 'A reason *for whom* to φ?', and we can further explicate these sentences as elliptical for sentences of the form, 'That *r* is a reason for *s* to φ.' (3) can naturally be interpreted as 'That plaque decays teeth is a reason [for you] to brush your teeth every day.' Accordingly, philosophers sometimes claim that the property of being a normative reason has argument-places for both an action φ and an agent *s*, hence with approximately the logical form *R*(*r*, *s*, φ).[2] A distinction is often drawn between "agent-relative" reasons and "agent-neutral" reasons, and a natural thought is that an agent-neutral reason is just a reason *for anyone* to φ.

[2] E.g. Schroeder 2007.

In seeking a unified logical form for reasons-talk, we may notice that 'for *s* to ψ' provides all the components of a proposition. So we may try reading 'a reason for *s* to ψ' as meaning *a reason that s ψ's*, just as (e.g.) 'I hope for *s* to ψ' seems equivalent to 'I hope that *s* ψ's'. A consideration in favor is that we do seem to ascribe normative reasons for the obtaining of propositions. Consider,

(5) 'The brutality of war is a reason for hostilities to cease.'

This has the grammatical form of an agent-relative reasons sentence, but it would be absurd to understand it as ascribing a normative reason to the agent, *hostilities*, to perform the action *ceasing*, rather than as reporting a normative reason favoring the state of affairs that hostilities cease; i.e. as equivalent to the sentence, 'The fact that war is brutal is a reason for its being the case that hostilities cease.'

Can we understand 'a reason for *s* to ψ' as *an explanation why s ψ's*? This seems impossible. First a grammatical difficulty: we can't simply substitute 'explanation why' for 'reason' in 'a reason for *s* to φ'. 'The brutality of war is an explanation why for hostilities to cease' is badly ungrammatical. More seriously, 'a reason for *s* to φ' seems to mean something quite different. A reason for hostilities to cease is not the same thing as an explanation why hostilities cease. One feature of talk about normative reasons is that it is in one way *nonfactive*: the existence of a normative reason for *s* to φ doesn't guarantee that *s* actually φ's. Yet 'explanation why' is factive: *r* is only an explanation why *p* if it is true that *p* (and *r*). Being explanatory of why *s* φ's is neither necessary nor sufficient for being a normative reason for *s* to φ. Counting in favor of something is evidently totally different from explaining it.

There is a natural way of translating normative reasons claims into explanatory claims, however. Plausibly, 'a reason to φ' is equivalent to 'an explanation why *to φ*'. The fact that plaque decays teeth is an explanation why to brush your teeth daily. If the fact that the subsequent train isn't until 11:30am is a reason to catch the 8:30am train, then it will be an explanation why to catch the 8:30am train. But although 'Why φ?' is a familiar kind of question, it requires interpretation, as 'φ' (e.g. 'Catch the 8:30am train') is not a valid explanandum. 'Why φ?' seems to be elliptical for some more complete question. But what?

We can presume an implicit agent of the action; Why: *I* catch the 8:30 train? Why: *you* brush your teeth daily? This provides the essential ingredients of a proposition, but we've already seen that

the question cannot be 'Why does *s* φ?' A natural alternative is to explicate 'Why φ?' as 'Why *ought s* to φ?'. 'A reason to brush your teeth' would thereby mean *an explanation why you ought to brush your teeth*. By 'a reason for *s* to φ', might we mean *an explanation why s ought to φ*, as some have proposed?[3] Instead of assigning 'reason' an essentially normative meaning, we thereby interpret it as meaning *explanation* in the ordinary sense, and locate the normativity in the explanandum. Since we are talking about normative reasons it makes sense that if they are answers to 'why?' questions then these will be *normative* 'why?' questions.

Although progressing from 'the reason to φ' to 'the reason *s* ought to φ' to 'the explanation why *s* ought to φ' may initially seem to preserve meaning, it can't be exactly right. The problem is that normative reasons often have merely pro tanto (contributory) rather than decisive "weight"; i.e. we can have reasons in favor of φ-ing without it being true that we ought to φ, if we have stronger reasons for not-φ-ing. Since 'explanation' is factive there can only be an explanation why *s* ought to φ if it is true that *s* ought to φ. So 'a reason for *s* to φ' cannot (always) mean 'an explanation why *s* ought to φ.' John Broome (2004) proposes that 'a reason for *s* to φ' is ambiguous between this and a pro tanto sense as *something that plays a role in a weighting explanation of why s ought or ought not to φ*. This solution is neither unifying nor compositional,[4] but there is an alternative which is.

In cases involving a defeated pro-tanto reason, calling it 'a reason *s* ought to φ' is no less unacceptable than calling it 'an explanation why *s* ought to φ'. So the difficulty here seems not to be that 'explanation why' is factive while 'reason' is not, but simply that we've identified the wrong kind of normative explanandum. If normative reasons only contribute a degree of weight toward what ought to be done, then a unifying analysis would have to identify a kind of explanandum that involves a normative concept which is contributory in this way. Here we may consider the widely accepted platitude that what ought to be done is whatever it would be *best* to do (which need not mean: whatever would yield the best consequences). Since to be "best" is just to be *most good*, *goodness* is a pro tanto normative concept that fits our needs. This suggests the hypothesis that 'a reason for *s* to φ' can be understood as elliptical for 'a reason [it would be good,] for *s* to φ', and means an explanation why it would be good, for *s* to φ.' The relation of *counting in favor* would

---

[3] Toulmin 1950: ch. 11, Finlay 2001: 104, Broome 2004: 34.
[4] See also discussion in Schroeder 2007: 36n, Brunero ms.

therefore be the relation of being explanatory of the goodness of. This is to reverse the popular "buck-passing" view that to be good is just to be favored by some reason.

This analysis overcomes our previous problems. First, while the construction 'an explanation why for *s* to φ' isn't grammatical, 'an explanation why [it would be good] for *s* to φ' is fine. Second, since the existence of a normative reason for φ-ing is plausibly sufficient for φ-ing's being good in some way, the analysis respects the factiveness of 'explanation': normative reasons explain *facts* by other *facts*. It also captures a relationship between reasons and value that many have thought correct.

Appealing to goodness also introduces some difficulties, as we shall see, but I will argue that these can be resolved by exploring two areas of vagueness or relativity in the analysis: in the notions (i) of being *good*, and (ii) of *explanation*. In particular, I will argue that they can be resolved by adopting an *end-relational* semantics for 'good'. On this theory (which I have championed elsewhere)[5] to say that *p* is 'good' is always to say that *p* is good for some particular end *e*, usually salient in the context. This is analyzed reductively as meaning that *p* promotes, or raises the probability, of *e*. This theory faces some obvious objections that I will not attempt to address here. The suggestion is then that to say '*r* is a reason for *s* to ψ' is to say that *r* is an explanation why it would be good for (some salient end) *e*, for *s* to ψ. This yields intuitively plausible results in mundane cases. The fact that plaque decays teeth is an explanation why it is good *for preserving your health* that you brush your teeth daily, just because it explains why brushing your teeth daily raises the probability that you preserve your health. Plausibly, the fact that the next train isn't until 11:30am may be an explanation why it is (e.g.) good *for getting to work on time* if you catch the 8:30am train, just because it may explain why catching the 8:30am train raises the probability that you make it to work on time.

3. Reasons for whom?

We commonly relativize reasons to agents with phrases of the forms, 'a reason *for s* to φ', '*s* has a reason to φ', '*s*'s reason to φ'. I have proposed reading 'a reason for *s* to φ' as elliptical for 'a reason [why it would be good] for *s* to φ'. This may suggest that a normative reason is relative to an

[5] Finlay 2006, ms. Analyses of normative reasons as promoting the agent's *desired* ends are rejected by Darwall 1983: 38, and defended by Schroeder 2007.

agent *s* just in case it favors a proposition *p* in which *s* is the "agent". But that can't be right. We cannot (e.g.) infer from (5) that hostilities "have a reason" to cease. Or suppose we say,

(6) 'The need to deter others from killing is a reason for murderers to be punished.'

Accepting this sentence doesn't incline us towards saying that murderers *have a reason* to be punished, or that the need to deter others from killing is the *murderer's reason* to be punished. "Having" a reason seems to involve having some special *normative* relationship to that reason. Instead, we might say that the need to deter others from killing is a reason *for us* (or for society, or judges, etc.), for murderers to be punished. Similarly, the brutality of war may be a reason *for statesmen*, for hostilities to cease. So it seems that accommodating the agent-relativity of reasons requires that 'a reason for *s* to φ' is semantically incomplete, and elliptical for 'a reason for s_1, for s_2 to φ'. It follows that 'a reason for *s* to φ' would be syntactically and logically ambiguous, between 'a reason [for s_1], for s_2 to φ' (e.g. 'a reason for hostilities to cease') and 'a reason for s_1, [for s_2] to φ' (e.g. 'a reason for you to catch the 8:30am train').

Plausibly, exactly the same kind of logical ambiguity is found in 'good for *s* to ψ' (compare 'It is good for murderers to be punished' with 'It is good for murderers to escape punishment'). So we can speculate that 'a reason for s_1, for s_2 to φ,' is elliptical for 'a reason [it would be good] for s_1, for s_2 to φ'. This may suggest that for an agent to *have* a reason to φ is for there to be an explanation why it would be *good for the agent* if she φ's. (On an end-relational theory of 'good', we can understand the qualifier 'for *s*' as indicating an end salient as being in *s*'s interest, so that 'good for *s*' is roughly equivalent to 'good for the obtaining of *s*'s interests'.) But that doesn't seem right, and we should reject this interpretation of 'a reason for *s*'. The first problem is that it tells us that our talk and thought about an agent's reasons to act always concern what is beneficial for him, or in his own interest. While ethical egoists maintain the controversial and fairly implausible thesis that all genuine reasons are from self-interest, it is totally implausible that this is what we ordinarily *mean* by 'a reason for *s*'. We contrast reasons of self-interest with reasons of altruism and self-sacrifice, for example. Second, while 'good for *s*' is plausibly equivalent to 'good for advancing *s*'s interests, [if *p*]', 'a reason for *s*' is not plausibly equivalent to 'a reason for advancing *s*'s interests', which can only mean 'a reason *to* promote *s*'s interests'. Finally, observe that in 'good for *s*, if *p*', *s* can be any object that can sensibly be ascribed an interest, including inanimate objects like trees and cars. But while 'good for

the car' is fine, 'a reason for the car' is not. Some cognitive powers seem required for "having reasons".

Equating reasons with *explanations* suggests an alternative account of the agent-relativity of reasons. If by 'a reason' we mean 'an explanation', then by 'a reason for *s*' we may mean 'an explanation for *s*', and by '*s* has a reason' we may simply mean '*s* has an explanation'. We do often relativize talk about explanations to subjects in this way—some fact *r* that is an explanation of *p* for one person s_1 may fail to be an explanation of *p* for another person s_2—and so we might reasonably expect 'reasons' to be relativized in the same way. This avoids all three problems just observed for the analysis in terms of 'good for *s*': (i) *r* can be an explanation for *s* of why it would be good in some way if *p*, without being an explanation why it would be good *for s*; (ii) an 'explanation for *s*' is not plausibly understood as meaning an explanation for some end *e*; explanations are for subjects, not for states of affairs; hence (iii) some cognitive powers are required for "having an explanation".

This predicts an ambiguity in '*r* is a reason for *s* to φ', between

(a) '*r* is an explanation for s_1, why [it would be good, for s_2] to φ,'
(b) '*r* is an explanation [for s_1, why it would be good,] for s_2 to φ'.

On the natural interpretation of (a), deletion of 'for s_2' suggests that s_2 is identical to s_1. So 'a reason for you, to catch the 8:30am train' is naturally read as 'an explanation for you, [of why it would be good, for *you*] to catch the 8:30am train.' On the natural interpretation of (b), the deletion of 'for s_1' suggests reference back to subjects salient in the context. So (e.g.) 'The need to deter others from killing is a reason for murderers to be punished' can be naturally read as (e.g.) 'The need to deter others from killing is an explanation [for us, of why it would be good,] for murderers to be punished.'

In order to determine whether the agent-relativity of reasons is properly analyzed in terms of the subject-relativity of explanations, we need to investigate exactly how explanations can be relative to subjects.

4. The Relativity of Explanations

In ordinary explanatory sentences like (1)-(2), the given explanans does not explain *all by itself*, but only in combination with other facts or information. With (1), the fact that the car didn't cross the sensor does not by itself reveal what makes it true that the light isn't turning green, but only against the background of information about the design of traffic signals, the weight of cars, etc. With (2), the fact that the prison fence is electrified does not by itself reveal what makes it true that the escape plan won't work, but only against the background of information that (e.g.) the plan calls for scaling the fence with bare hands, etc. Some therefore claim that (e.g.) the fact that the prison fence is electrified is not *really* the explanation why the escape plan won't work, but merely *part of* the explanation.

The same issues arise for talk about normative reasons. The fact that (e.g.) the subsequent train isn't until 11:30am doesn't count in favor of catching the 8:30am train all by itself, but only against the background of other facts, such as that it takes you to your office and that you need to be in your office by 9:30am. This is the "holism of reasons": what counts as a reason to φ in one context might not do so in other contexts (e.g. in which the train doesn't take you to your office), just like what counts as an explanation of *p* in one context might not in other contexts. Some therefore argue similarly that the fact that (e.g.) the next train isn't until 11:30am is not really the reason for catching the 8:30am train, but merely part of the reason. I don't need to rule either way on this debate, but will continue in the ordinary way of speaking.

What we *report* as reasons or explanations why are the facts which, against a body of assumed information *B*, are decisive in revealing what makes something the case. Hence, explanations are relative to bodies of background information *B*. In one clear sense, to be an explanation *for s* is to be an explanation *relative to the information B that s possesses*. The fact that the prison fence is electrified may be an explanation why the escape plan won't work for a person who knows the content of the plan, but it is not an explanation why the escape plan won't work for someone who doesn't have that information.[6] The fact that the subsequent train isn't until 11:30am is an explanation for you why it would be good (for your getting to work on time) for you to catch the 8:30am train, *if* you already know that the train takes you to the office, but not otherwise. We can therefore try

[6] Although having it presented *as* an explanation may enable her to deduce the plan's contents.

analyzing 'a reason for *s* to φ' as meaning an explanation, relative to *s*'s information *B*, why it would be good, for *s* to φ. This suggests identifying a covert argument-place in the logical form of reasons-statements for an information-base *B*: we have the form *R*(*r, p*, *B*), meaning that *r* explains *p* relative to *B*.

While this information-relativity accounts for some behavior of 'a reason for *s*', it does not seem the principal way normative reasons are agent-relative. We need a kind of *normative* rather than epistemic connection between agents and their reasons. That the next train isn't until 11:30am may also be an explanation *for me*, of why it would be good for your getting to work on time, for you to catch the 8:30am train. But this wouldn't automatically make it a reason *for me*, as well as a reason for you, [for you] to catch the 8:30am train. (I might lack any reason favoring your getting to work on time). Neither is this difference explained merely by the fact that it is a reason why it would be good if *you* φ, and not if *I* φ, as the following case shows. Let *r* be the fact that Terry has stolen Victor's wallet. Relative to Terry's and Victor's information, that *r* is an explanation why it would be good (for Victor's recovering his wallet), for Terry to be apprehended. But while we can say that *r* is a reason for *Victor*, for Terry to be apprehended, we would not say that *r* is a reason for *Terry*, to be apprehended. So 'a reason for *s*, to φ' cannot simply mean 'an explanation for *s*, [why it would be good (in some way), for *s*] to φ'.

A better solution emerges when we return our attention to the relativity in 'good'. According to an end-relational account of 'good', 'a reason for *s* to φ' will be ambiguous between 'an explanation why it would be good *for* e_1, for *s* to φ' and 'an explanation why it would be good *for* e_2, for *s* to φ', etc. So we may speculate that qualifying a normative reason as being 'for *s*' functions to indicate the intended kind of end-relative goodness in the explanandum. By 'a reason for Terry' we would thereby talk about an explanation why something would be good for an end *e* made salient somehow by reference to Terry. Since the agent-relativity of reasons is a *normative* connection it makes sense that it would arise from the normative part of the analysis.

How does reference to an agent make a particular end *e* salient, if not by directing us to the agent's welfare/ what is good for *s*? The following account is a natural alternative. If a reason or explanation is an answer to a potential "why?" question, then we might understand talk of 'an explanation for *s*' as meaning an answer to a "why?" question made salient by reference to *s*. Plausibly, this would be a question 'Why would it be good to ψ?' that *s* might himself ask in

deliberating about what to do: a question framed by ends *that matter to s*. We can naturally say (e.g.) that the fact that Terry has just stolen Victor's wallet is an explanation *for Victor* of why it would be good, for Terry to be apprehended, but it is not an explanation *for Terry* of why it would be good, for him to be apprehended. Here the agent-relativity of explanations is explained not by sensitivity to information, but by the different readings of 'good' made salient by the difference in Victor's and Terry's concerns; that is,

> 'That *r* is an explanation for Victor$_i$, of why it would be good [for his$_i$ ends], for Terry to be apprehended.'
>
> 'That *r* is an explanation for Terry$_i$, of why it would be good [for his$_i$ ends], for Terry to be apprehended.'

This leads us towards a Humean theory of an agent's reasons, as those considerations relevant to the satisfaction of his desires. But while a Humean theory seems to account for many ordinary uses of 'a reason for *s*', the analysis allows that it may be too narrow, and that we may ascribe normative reasons for an agent *s* to φ that are explanations why *s*'s φ-ing would be good relative to ends that are salient in other ways, e.g. for being of interest to other agents. This is just as well, since ordinary talk about normative reasons does not always respect the limits imposed by Humeanism.

5. Reasons for Attitudes

A difficulty arises when we try to extend this analysis of talk about normative reasons as explanations why to accommodate also our talk about normative reasons for *attitudes*, such as belief, intention, desire, fear, remorse, etc. Consider

> (7) 'Rosemary's performance tonight is a reason for her to be ashamed.'

Can we analyze this sentence in the way suggested above, as equivalent to 'Rosemary's performance tonight is an explanation for why it would be good, for her to be ashamed.'? Having certain attitudes can be beneficial. Emotions are effective motivators, and often dispose agents in desirable ways. Shame might motivate Rosemary not to inflict her talentless act on any future victims, or to practice more. Intentions dispose people to act, so if φ-ing would be a good thing, then in general *intending* to φ would also be good—hence a reason to φ will in general be also a reason to intend to φ.

The analysis thus easily accommodates "state-given" reasons for attitudes, based on the benefits of having those attitudes. But ordinary talk about reasons for attitudes is concerned with "object-given" reasons, based not on any benefits of those attitudes but on the nature of those attitudes' *objects*. For example, (7) is naturally read as meaning that Rosemary's performance is a reason for her to be ashamed *of her performance*, where this reason exists because shame is an appropriate or "fitting" attitude for Rosemary to have towards her performance, because it was shameful. The fittingness of an attitude is typically independent of the attitude's consequences. The quality of Rosemary's performance may give her a reason to be ashamed even if being ashamed promotes no positive outcomes, only negative ones like losing all self-esteem, abandoning her art, and the grief of her friends. This presents a challenge to analyzing these normative reasons as explanations of goodness.

If our semantics for 'reason' is to be unifying, it must also provide a plausible analysis of claims about reasons for *belief*. Since reasons for belief are a kind of normative reason, the analysis suggests that by 'a reason to believe that *p*' we mean an explanation why it would be good, for *s* to believe that *p*. But instrumentalist theories in epistemology don't seem plausible. Even if there are "pragmatic" reasons for belief (e.g. that believing will make you happy), this is not how we ordinarily understand 'a reason to believe that *p*'. *Epistemic* reasons for *s* to believe that *p* consist in *s*'s evidence that *p*, and hence are object-, not state-given. Evidence that *p* makes belief that *p* fitting, but doesn't require that it would be good for any end *s* desires.

However, the theory doesn't require us to give 'a reason for *s* to φ' a narrowly instrumentalist interpretation, as an explanation why *s*'s φ-ing would be good *for some end desired by s*; it requires only that there be some salient way of being good. Might 'a reason to believe that *p*' mean an explanation why believing that *p* would be *epistemically* good? What we need is that 'a reason to believe that *p*' makes salient some kind of epistemic end. Plausibly it does, since belief is widely considered an attitude with its own "constitutive aim": an end at which an attitude must aim in order to be a belief. (Insofar as talk of reasons is intended to influence people, our inability to believe for any other kind of reason will force the salience of the constitutive end). In a slogan, belief aims at *truth*. Without trying to settle disputes about to develop this idea, we can now show that a standard kind of model of the aim of belief supports a plausible solution.

Cognitive activity of forming a belief that *p* aims, roughly, at thereby believing that *p if and only if p is true*. So we can speculate that 'a reason for *s* to believe that *p*' is elliptical for 'a reason for *s*, [why it would be good for thereby believing that *p* iff *p* is true, for *s*] to believe that *p*', and is reductively analyzable as *an explanation for s, of why it would increase the probability that s thereby believes that p iff p is true, if s believes that p.*[7] So if I say 'The fact that the prison fence is electrified is a reason for Arthur to believe that the escape plan will fail,' I mean that this fact is an explanation for Arthur of why his believing that the plan will fail would increase the probability that he thereby believes that it will fail iff it is true. Applying the analysis of 'an explanation for *s*', this is to say that the fact that the prison fence is electrified is something that reveals to Arthur, in light of his other information *B*, what makes it true that his believing that the escape plan will fail would increase the probability that he thereby believes the plan will fail iff that is true. If I am right that this yields a fair approximation to the truth-conditions for '*r* is a reason to believe that *p*', then this semantics for 'reason' can be extended to talk about epistemic reasons too.[8]

We can treat this as a template for analyzing talk about reasons for other kinds of attitudes. Here's a quick sketch of the idea. Plausibly, any kind of attitude which is made fitting by the nature of its objects is one that has a characteristic kind of constitutive aim. In the same way that belief aims at the true, shame aims at the shameful and fear aims at the fearsome. This suggests analyzing Rosemary's reason to be ashamed, for example, as an explanation why it would good for Rosemary's thereby being ashamed of her performance iff it was shameful, for her to be ashamed of her performance. A reason to be afraid of *x* would be an explanation why it would be good for thereby being afraid of *x* iff *x* is fearsome, to be afraid of *x*. These analyses seem defensible, even if not very informative. I speculate that these emotions can be reductively analyzed as a complex of cognitive and conative elements, so that (e.g.) to be ashamed of *x* is to be *averse to attention to x as being to one's discredit*. Hence, Rosemary's reason is an explanation why it would be good for Rosemary's thereby being averse to attention to her performance as being to her discredit iff her performance was to her discredit, for her to be ashamed of her performance.

---

[7] "*Thereby* believes" because the fact that by *falsely* believing that *p* one can make it so that in the future one correctly disbelieves that *p* is not an epistemic reason to believe that *p*: the salient epistemic end in believing that *p* is concerned only with that very token of belief. A parallel solution applies for deontological side-constraints in morality: e.g. you may not kill even to prevent more killings.

[8] See Chrisman 2008 for an end-relational account of 'ought to believe'.

More work is needed here, but if I am right that these analyses are promising then the analysis of 'reason' as meaning *explanation why* can be extended in a systematic way to our talk about reasons for attitudes, by taking the reference to an attitude as making salient a constitutive end *e*.

6. The Reasons for Which We Act

Superficially, talk about "motivating" reasons for action wouldn't seem to present any obstacle, since attributing a motivating reason to an agent is a way of explaining action. We might try viewing this simply as a case of explanatory reasons where the explanandum is that *s* φ-ed. On an orthodox theory, motivating reasons for action are explanations why *s* φ-ed that cite the psychological attitudes (beliefs and desires) that caused *s*'s φ-ing;[9] for example,

(8) 'The reason why Cleopatra had herself rolled up in a carpet was that she wanted an audience with Caesar and believed that this would get her one.'

However, this simple strategy encounters difficulties, and I believe it is confused.[10]

Consider these ordinary motivating reasons sentences;

(9) 'The reason for which Cleopatra had herself rolled up in a carpet was that it would get her an audience with Caesar.'

(10) 'Cleopatra's reason for having herself rolled up in a carpet was that it would get her an audience with Caesar.'

These sentences seem to identify Cleopatra's reason as the fact *that r*: having herself rolled up in a carpet would get her an audience with Caesar—and not any facts about her beliefs and desires. We should also observe the difference between 'the reason *why s* φ-ed' and 'the reason *for which* s φ-ed';[11] we don't find the latter locution in other non-normative 'reason' sentences—e.g. we wouldn't say that the fact that the prison fence is electrified is the reason *for which* the escape plan isn't going to work. It would also be peculiar to say that the reason "for which" Cleopatra had herself rolled up in a carpet was that she wanted an audience with Caesar and believed this would get her one. It seems,

[9] Views of roughly this kind are advanced in Davidson 1963, Smith 1994.
[10] My objections here largely follow Darwall 1983, Dancy 2000, Alvarez 2010.
[11] Williams 1979.

therefore, that the causal-psychological reasons why an agent acts cannot be what we mean by 'the reasons for which she acts'.

Notice however that the fact *that r* which (9) and (10) report as Cleopatra's motivating reason is itself a candidate for being a *normative* reason for her to φ. The fact *that having herself rolled up in a carpet would get her an audience with Caesar* was an explanation, for Cleopatra, of why doing so would be good for some further end of hers (like winning Caesar's favor). This suggests a different solution: by '*s*'s reason for φ-ing' and 'the reason for which *s* φ-ed', we mean *the reason to φ for which s φ-ed*. To be a motivating reason is to be a (normative) reason that motivates, not simply a reason why the agent is motivated. We could then understand (9) as elliptical for

> (9') 'The reason [it would be good (for some end of hers), for her to have herself rolled up in a carpet,] for which Cleopatra had herself rolled up in a carpet, was that it would get her an audience with Caesar.'

Normative reasons are things agents are supposed to take into consideration in deliberating about what to do. Surely then, when all is going well agents are motivated as a result of their awareness of their normative reasons. Plausibly, this is what we mean in saying that *r* is the reason "for which" *s* φ-ed: *r* is what *s* accepted as a reason to φ, leading her to φ. By contrast, the rival analysis of motivating reasons—as facts about an agent's psychological attitudes (or as those attitudes themselves) that causally explain her actions—results in mismatch between normative and motivating reasons, because facts about her psychological states are not typically among the normative reasons to which an agent attends and responds in deliberation.

Although claims about the reasons that motivate *s* to φ do explain why *s* φ's, confusingly this is not what we mean in calling them 'reasons'. Rather, these claims explain actions by revealing what the agent accepts as her normative reason, or explanation why it would be good for her to φ, which motivates her to φ. This reduction of motivating reasons to normative reasons for which agents act also encounters an obstacle, however. Arguably, *whenever* agents act, they act for reasons. It's natural to say, 'She *must* have had a reason,' and we would be suspicious of anyone who claimed to have intentionally done something for no reason at all. But sometimes people do things which we would not say there was any (normative) reason to do, things we would deny were "good" in any salient way. We can distinguish between two different kinds of case.

The first case is where an agent is motivated to φ by a fact *that r* that isn't really a normative reason to φ, because it doesn't count in favor of φ-ing. Suppose Sarah drives to her daughter's school (φ) for the "reason" that the time is 3:30pm (*p*), which she takes to explain why φ-ing would be good for picking up her daughter on time (*e*)—but has momentarily overlooked that her daughter is going home with a friend that day. There isn't any reason for Sarah to φ; in this case *s* is mistaken in taking *that r* to explain why her φ-ing would be good for *e*. But while here we rightly deny that there is any reason for *s* to φ, it is still right to say that *s* φ's "for a reason"; Sarah's reason for driving to school is that the time is 3:30pm. It seems a fact can be a motivating reason for φ-ing without being a normative reason for φ-ing.

The second case is where an agent is motivated to φ by taking *that r* to explain why φ-ing would be good for *e*, but is mistaken in believing that *r*. Consider Bernard Williams' agent *s* who drinks from a bottle (φ) containing petrol, because he mistakenly believes that it contains gin (*r*). If you were to ask *s* for the reason for which he is about to drink from the bottle, he may respond, 'I'm going to drink from the bottle for the reason that it contains gin'; i.e. he may identify his own motivating reason as the fact that the bottle contains gin. But this "fact" *that r* does not exist. While there is no reason for *s* to φ, we would still say that *s* φ's "for a reason". It seems we can have motivating reasons for φ-ing when there are no facts that are candidates to be normative reasons for φ-ing.

One response is to say that agents are sometimes motivated by normative reasons *that don't exist*. This is not as absurd as it may sound, given that we understand being motivated to φ by a normative reason *that r* as a matter of being motivated to φ through *accepting that r* as a normative reason to φ. Since in this explanation *that r* occurs in an intensional context, it doesn't require that *r* obtains, just as our being scared by Count Dracula and saddened by the death of Bambi's mother don't require those things to be actual. A problem remains for this solution, however; we still say (e.g.) 'There is a reason for which Sarah drove to school', which is apparently an *existential* claim. How can there be a reason when there isn't a reason? How could *that r* be *s*'s reason, when there is no such reason as *that r*?

A problematic piece of data here is that in the second case, when a speaker realizes that *r* does not actually obtain it is natural for her to describe the motivating reason differently; *after* taking a sip from the bottle, Williams' agent would rather say, 'I drank from the bottle for the reason that *I*

believed it contained gin', or 'My reason was that *I believed* it contained gin'. These are answers that point us back towards the first analysis of motivating reasons as the psychological facts or attitudes that causally explain the action. There is thus an asymmetry in how speakers describe motivating reasons depending on whether they accept or reject the agent's belief. Unless we say that agents are motivated differently between the illusory and veridical cases, it seems that one or other way of describing motivating reasons must be misleading. But which?

Michael Smith argues that it is the description in the veridical case that is misleading; strictly, an agent's belief is always part of her motivating reason. But the philosophical considerations above seem to me decisive: the reasons for which agents act must be the considerations to which they attend and respond in their deliberations, and these do not typically include facts about their attitudes. Furthermore, we can explain the interposition of '*s* believed' in the illusory case: saying '*s*'s reason for φ-ing was that *r*' suggests that the speaker accepts that *r*, a suggestion that can be blocked by the interposition of '*s* believed'—or more transparently, as Dancy proposes, by '*as s* believed'. Williams' agent seems more accurately to say, 'I drank from the bottle for the reason that, as I believed, it contained gin.' I therefore suggest that saying 'I drank from the bottle for the reason that it contained gin' is indeed not to say something false, but merely misleading.

The need to include such a disclaimer in order not to mislead is however a piece of data that leads us back to the main problem for the proposed analysis: that it interprets these sentences as claiming the existence of normative reasons that the speakers know do not actually exist. One solution here is to give up on a unifying semantics, and distinguish between *objective* and *subjective* normative senses of 'reason', where an agent *s*'s subjective reason to φ is defined as a *proposition* that *s* takes to be an objective reason for her to φ.[12] We can then disambiguate, and say that whenever an agent φ's there must be a subjective reason, or proposition that she takes to be a reason to φ, for which she φ's, even if she is mistaken and there is no such objective reason to φ.

But we should want to say, of the veridical case in which *s* recognizes the fact *that r* as a reason to φ, that she is motivated through her awareness of the objective reason she has to φ, and that the reason for which she φ's is her objective normative reason, the *fact that r*. So if we are to avoid treating motivation in the veridical and illusory cases differently, we should say in the illusory case that she takes as her (objective) reason to φ the supposed but nonexistent *fact that r*, and that she

[12] E.g. Darwall 1983: 32, Schroeder 2007.

is motivated through her acceptance of this supposed reason *that r*. The hypothesis that by 'reason' we always mean *explanation why* enables us to sweeten this pill. Notice that while 'the explanation why' is factive, '*s*'s explanation why' behaves differently. If I say (e.g.) 'Arthur's explanation for why the escape plan won't work is that the prison fence is electrified', I haven't necessarily implied that the escape plan indeed won't work, or that the prison fence is indeed electrified, or even that the one would indeed explain the other. '*s*'s explanation why…' is therefore distinct in meaning from 'the explanation for *s* why…' Rather, by 'Arthur's explanation' I seem to mean approximately 'that which Arthur *takes* to be the explanation'. Here scare-quotes seem appropriate to flag the spuriousness of this supposed explanation (and, I suggest, perform the same function as '*s* believes')[13]: 'Arthur's "explanation" for why the escape plan won't work is that the prison fence is electrified.' This matches what we've found in the case of motivation by supposed reasons;

> 'The "explanation" [*s* accepted for why φ-ing would be good,] for which *s* φ-ed, was *that r*.'

This enables us to evade the apparent contradiction to which our reasoning had led us. *s* can have an "explanation" why φ-ing would be good, and be motivated by that "explanation", even though there is no actual explanation why φ-ing would be good, either because *that r* doesn't genuinely explain this, or because there is actually no fact *that r*.

Does this solution preserve the semantic unity of 'reason', however, or does it merely disguise an ambiguity by employing another word concealing the same ambiguity? We might avoid that conclusion by understanding these inverted commas sentences as follows:

> 'The [supposed fact *s* took to be a] reason [why drinking from the bottle would be good,] for which *s* drank from the bottle, was that it contained gin.'

Or merely, 'The [supposed] reason for which *s* φ-ed, was that *r*'. This extends my unifying semantics to accommodate talk about the 'reasons for which *s* φ's'.

---

[13] Inserting '*s* believes' may result from an understandable confusion about these sentences. Because we know (i) that 'reason' means *explanation why*, (ii) that these sentences purport to explain actions, and (iii) that explanations are factive, we are reluctant to say that the reason was *that r* when we don't believe *that r*, and we turn to the fact *that s believes that r* as an acceptable substitute, being also an explanation why *s* φ's. This involves failing to recognize that in using these sentences we are explaining *s*'s φ-ing by reporting what *s* took to explain why φ-ing would be good.

7. Conclusion

I have argued that a semantic theory of 'reason' as meaning *explanation why* can give unifying and reductive analyses of our talk about explanatory reasons, normative reasons for actions and attitudes, and motivating reasons. On this analysis, by a normative 'reason to act' we mean an explanation why it would be good, in some salient way, to act. This analysis of normative reasons sentences seems to yield correct truth conditions for ordinary cases if we adopt a reductive analysis of 'good' as end-relational.

References

Alvarez, M. (2010), *Kinds of Reasons*. Oxford: Oxford University Press.

Broome, J. (2004), 'Reasons', in R. J. Wallace et al. (eds.), *Reason and Value*. Oxford: Oxford University Press, 28-55.

Brunero, J. (ms), 'Reasons as Explanations'.

Chrisman, M. (2008), 'Ought to Believe', *Journal of Philosophy* 105: 346-70.

Dancy, J. (2000), *Practical Reality*. Oxford: Oxford University Press.

Darwall, S. (1983) *Impartial Reason*. Ithaca, NY: Cornell University Press.

Davidson, D. (1963), 'Actions, Reasons, and Causes', reprinted in *Essays on Actions and Events*. Oxford: Oxford University Press, 1980: 3-19.

Finlay, S. (2001), *What Does Value Matter?* University of Illinois Ph.D. Dissertation.

(2006), 'The Reasons that Matter', *Australasian Journal of Philosophy* 84: 1-20.

(ms), *Confusion of Tongues*.

Schroeder, M. (2007), *Slaves of the Passions*. Oxford: Oxford University Press.

Smith, M. (1994), *The Moral Problem*. Malden, MA: Blackwell.

Toulmin, S. (1950), *Reason in Ethics*. Cambridge: Cambridge University Press.

Williams, B. (1979), 'Internal and External Reasons', in R. Harrison (ed.) *Rational Action*. Cambridge: Cambridge University Press.