

## Explanation, Justification, and Motivating Reasons

At the heart of many accounts of agency lies the idea that we can act for reasons. When we act for reasons, those reasons are supposed to both explain and to justify our actions. What we commonly say in such instances is that we acted *because* we had certain reasons, intending the relevant *because* to capture an explanatory as well as justificatory relationship between our reasons and our actions. The kind of reasons in question are *explanatory* or *motivating reasons*, and need not be *good* or *objectively justifying reasons* – in other words need not be *normative reasons*. Still, as Michael Smith writes, what explanatory and normative reasons “have in common is that each purports to justify certain behavior on [an agent’s] behalf.”<sup>1</sup> With respect to motivating reasons, one plausibly adds the qualification “justifies from the agent’s perspective,” or even, as Smith suggests, from the “perspective of the value that that very [motivating] reason embodies.”<sup>2</sup> Notice, however, that such qualifications only restrict, but do not revoke, the justificatory aspirations of motivating reasons altogether. Even if restricted to perspectives of values and agents, motivating reasons purport to say something in favor of certain actions. Thus, Derek Parfit is correct when he says “*motivating* reasons can thus be regarded *both* as normative reasons *and* as motivating states.”<sup>3</sup>

I wish to raise some doubts as to whether there can be such things as explanatory reasons as so construed (*ERs* henceforth). I do believe we have decent accounts of how there could be states that explain in virtue of some of their features while justify in virtue of others. But I think we are still at a loss in understanding how states could explain *in virtue of having justificatory*

---

<sup>1</sup> Michael Smith, *The Humean Theory of Motivation*, Mind 96 (1987), p.38.

<sup>2</sup> *Ibid.* p.39.

<sup>3</sup> Derek Parfit, *Reasons and Motivation*, The Aristotelian Society, Supplementary Volume 77 (1997), p.114 (footnote)

*features*. The difference between justification being merely conjoined with explanation versus justification being truly involved in explanation is subtle, but crucial. Consider Venice, which is both beautiful and in danger of being inundated in the ocean, but which is not in such danger in virtue of being beautiful. Aesthetic properties do not literally have inundating powers. Likewise, ERs may explain in virtue of having certain features and justify in virtue of having other features. But even if this is true, it falls short of making good the idea that there is one unified *because* involved “A does  $\phi$  because she has reason to  $\phi$ .” If the justificatory potential of reasons is to enter that *because*, it must be so because it makes an explanatory contribution rather than just dragging along. It is easy to see how justification could quasi epiphenomenally accompany explanation. It is not so easy to see how justification could truly be involved in explanation. My thesis will be that reasons qua justifiers do not explain, and qua explainers do not justify.

Let me begin by focusing more closely on what is involved in explanation and justification. I will follow Davidson in assuming explanation to be ‘causal explanation.’ While one may choose to use ‘explanation’ in alternative ways, the causal sense has certainly dominated the literature, and in this short essay I will restrict my attention accordingly. Thus understood, explanation and justification apparently present two rather different relationships. If A explains B, A is a causally relevant factor in bringing about B. If A justifies B, B becomes normatively appropriate in light of A. Specifying more precisely what the difference consists in presents a philosophically challenging task, but I believe one can best approach the matter by considering how we evaluate “unusual” cases. Suppose that usually A-states bring causally about B states, suppose for instance that economic theory is correct in postulating an explanatory relationship between the Feds lowering the interest rates and an increase of inflation. Now suppose further that this time we find no increase in inflation to follow a decrease in interest

rates. How do we evaluate such departure from common causal patterns? Well, we would say it is unusual, or uncommon, or perhaps surprising. We didn't expect that coming. But it would be quite a stretch to say that it was somehow inappropriate for B not to occur given that A occurred, or that it was inappropriate for inflation not to go up given that interest rates went down. Yet such a verdict would be characteristic if we were dealing with a justificatory relationship. If A justifies B, then – *ceteris paribus*, and from the agent's perspective – B's absence given A must not be seen as merely unusual, but as a failure of sorts, as something that should have happened but didn't. For if A justifies B, then only the scenario *B in light of A* but not the scenario *non-B in light of A* accords with how things were supposed to turn out. Under certain conditions, and in case it was a person who was responsible for the absence of B given A, that person may come to be subjected to criticism, whereas no critical force can spring from merely breaking causal patterns.

Now keeping this important difference in mind, let me present my analysis of why ERs cannot explain in virtue of having certain justificatory features. My argument will proceed by addressing the best case I'm aware of for construing ERs as explainers in virtue of being justifiers, which introduces a classical representational model. Suppose the reason why Peter donated money to famine relief was that he thought it was the right thing for him to do. Obviously, if the reason for his donation was (merely) a temporary hormonal increase or something of that sort, we would be off a bad start in accounting for the justificatory potential of his explanatory reasons. So let's just grant that (1) it was indeed the right thing for him to donate money, (2) that he was aware of that fact, and (3) that his awareness plays an explanatory role for why he donated money. In this model, Peter's ER would consist in some motivationally

efficacious representations of a (normative) reason he has to donate money. Why is there still a problem with Peter acting as he did *because* he had a justificatory reason to donate money?

To see why, notice first that given our representational model, what justifies and what explains are distinct entities. What justifies are the contents of representational states (or their relationships), and what explains are causal properties of content-bearers (or vehicles). What justifies Peter's behavior is what he was aware *of*, namely the fact that donating money is the right thing for him to do. What set his behavior into motion was the state of awareness, or the representational state which has such and such content. This by itself doesn't prove much, but nonetheless provides starting point.

To see where this is leading, consider belief first. Mary believes  $p$  and *if  $p$  then  $q$* , and comes to believe  $q$ . Now, what justifies this transition in belief are the contents of those beliefs together with the logical relationship of entailment that holds between those contents. The content of her (initial) beliefs are the propositions  $p$  and *if  $p$  then  $q$*  and those proposition entail the proposition  $q$ , which of course is the content of the belief she arrives at. Yet propositions and especially logical relations between propositions are abstract objects and causally inert. Since representational states but not logical relations between contents have causal powers, what caused Mary's transition in belief must be representational states rather than relations between contents. The fact that the propositions  $p$  and *if  $p$  then  $q$*  entail the proposition  $q$  cannot be what made it the case that one representational state causally brought about another.

Certainly the transition between the representational states can match or mirror logical relationships between contents. Ever since logical theory has managed to establish syntactical systems exhibiting deductive behavior which correlates to semantic soundness relations between

propositions, this has inspired many defenders of representational systems to try explaining how systems can behave for reasons. The idea is to postulate two levels that neatly match, where one (the level of vehicles) does the causal work and the other (the level of contents) does the justificatory work. But since correlation is not the same as causation, this picture still doesn't make good the idea that representational states can causally bring about other representational states in virtue of having content relationships that satisfy logical relations. All the transitions between representational states seem in principle explainable solely by attending to how they interact in given representational systems.

For illustration, consider two systems. One delivers output  $A$  given the input  $(A \& B)$ , the other delivers  $A$  given  $(A \text{ OR } B)$ . Suppose both systems have an array of buttons one can press (the input, "A", "B", "C" ... plus "&", "OR", etc.), and a tape which then prints other symbols (output). Let us grant that the letter symbols represent propositions and the logical symbols represent logical functions. Now, when we try to explain why one system behaves one way whereas the other system behaves another way, will the fact that only one system exerts valid inference behavior while the other does not make any causally explanatory contribution? It doesn't seem so. Rather, what would explain such inference-behavior would be the functioning of the relevant representational mechanisms. We would take them apart, and try to figure out how representational states interact in those systems.

Now, just as the representational model cannot make good the idea that belief-vehicles cause in virtue of representing contents exhibiting certain logical relations to each other, so I would like to argue the representational model can't make good the idea that ERs cause in virtue of having justificatory features either. But let me pause for a moment and say why in practical contexts we encounter even graver complications than we encountered in theoretical context

such as belief. I have taken it for granted that the relata of justificatory relationships are contents. Yet while in theoretical contexts one may cite truth or implication-relationships between contents as grounding justificatory relations (*A&B justifies A because A&B implies A*), it is anything but clear in virtue of what certain practical contents can justify others. The most prominent candidates – belief/desire pairs – do not entail intentions or descriptions of actions, not least because desires and (probably) intentions don't have truth conditions (recall all the futile attempts of expressivists to deal with the Frege-Geach problem). So whatever accounts for such justificatory relations, it can't simply be their contents alone, but must introduce some heavy duty practical principle. Yet apart from difficulties arising for the justification of such a principle, it is not even clear how to formulate it yet. Any plausible candidate I'm aware of brings with it an unmanageable array of *ceteris paribus* clauses. The simple practical syllogism which says that 'if one desires *q* and believes *q if p* one ought to/should/is rationally permitted to do *p*' is certainly invalid as it stands. One ought not to do *p*, for instance, if doing so conflicts with other more substantial aims, or if one's belief *if p then q* is false, etc.

Now, let me return to my analysis of why ERs cannot explain in virtue of justifying. Explanations broadly construed answer why questions, and there are at least two different kinds of why-questions that may be at stake here. On a local level, the question may concern why or how some particular state was brought about given some other starting states obtained. On a more global level, the question may concern why certain systems at large function the way they do. Consequently, my claim that no explanatory why-question will be answered by referring to justificatory facts can be assessed on a local or global level. Let's turn to local explanation first.

My argument here will be similar to what has already been sketched with regard to belief. In a nutshell, I see a fundamental tension between the idea of explanation in virtue of

justification and an increasingly well supported naturalistic conception of our minds and psychology. What I seek to bring out is a misfit between the idea that representational states explain in virtue of having justificatory features and with what the Nobel Prize winning biologist, Francis Crick, bluntly described as "The Astonishing Hypothesis," the hypothesis that "You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules. As Lewis Carroll's Alice might have phrased it, "you're nothing but a pack of neurons."<sup>4</sup> If all the work in explaining certain representational transitions is done by physical features of those representational states together with features of the representational systems, what is the explanatory contribution that justificatory facts are supposed to add? As far as we can see, what explains the transition from one particular representation to another is that the relevant content-vehicles bear some causal relationship to each other within some representational system. One could say that given that *this* is how the relevant representational system works, *this* content-vehicle will cause *that* content-vehicle. What explains their causal relationship is not some justificatory relationship between contents, but rather the fact that this is the way content-vehicles function in some given system.

Now, more globally directed explanation will not imbue justificatory relationships with explanatory potential either, though I grant that the global case seems more promising than the local case. In the context of global explanation, the task is to explain why our representational systems at large function the way they do. We have a tendency to deduce  $A$  from  $(A \& B)$  but not from  $(A \vee B)$  (though there's evidence for being more pessimistic. For instance, we also have a

---

<sup>4</sup> Quoted from Paul Bloom, Introduction to Psychology, Yale online courses, transcript here: <http://open.yale.edu/courses/psychology/introduction-to-psychology/transcripts/transcript02.html>

tendency to deduce *not-B* from  $(A \supset B)$  and *not-A*, which is invalid), and one may suspect that this has something to do with the fact that one inference but not the other is valid (hence would mention justificatory aspects). Yet again, I think this is false, though admit that I will not be able to adequately address this largely empirical issue here. As far as I'm aware of the evidence, it is unlikely that justificatory facts will enter the causal explanation of why our representational systems function one way rather than another. Rather, I believe that facts concerning our selective and/or developmental history in which our systems evolved will be key. Given the complex context of our evolutionary history there must have been some differential selective advantages accruing to certain design patterns which accumulatively generated our representational systems. In principle, though not in practice, one may point out at any stage of this history why on balance systems with certain features reproduced in greater numbers than systems with other features. From an explanatory standpoint, and to put it somewhat crudely, the fact that the desire not to be bitten by angry dogs and the belief that one can avoid that result by staying away from angry dogs usually produces the intention to be away from angry dogs is an accident of our history. Alter certain features, and the aforementioned belief/desire pair might have systematically produced the desire to get close to angry dogs. Explanation per se is indifferent to whether our ways of processing contents follows justificatory prescriptions. Again, it might, but even then it is unclear and I think doubtful whether our ways of processing contents which accords with certain prescriptions behave that way *because* they accord with those prescriptions. The fact that there's a justificatory relationship between those contents per se does not explain why we ended up with a representational system which behaves one way rather than another, but at best only points us to look in certain directions.

Ultimately, the issue may be empirical, and not up for philosophers to decide. For many philosophers, this would be no minor concession. Their confidence in the dual powers of explanatory reasons does not appear to spring from any particular empirical evidence, but rather from metaphysical assumptions about agency. If it can be shown that the matter of dual powers of ERs rests on empirical findings, I believe my challenge would have achieved all a philosopher could possibly wish for.

Let me close by making one final point. On all accounts, theory development of ERs as explainers qua justifiers has encountered severe obstacles. Perhaps we must try harder. But perhaps my hypothesis that justificatory and explanatory potential at best springs from different aspects of ERs affords us some insight about why such grave difficulties have affected traditional theory development of ERs in terms of explanation in virtue of justification. After all, making a theory of ERs adequate as a theory of justification gives rise to rather different constraints than making it adequate as a theory of explanation. Qua theory of explanation, the theory's merit is a function of whether it correctly reconstructs the causal processes which brought about some piece of behavior. Qua theory of justification, the theory's merit is a function of whether it can correctly account for why certain states justify other states. These constitute two distinct merit functions, and trying to satisfy both functions simultaneously presents some fundamental problems. We may thus be better off divorcing theory development of motivating reasons, and instead of advancing one single theory of a heterogeneous phenomenon focus either exclusively on explanation or exclusively on justification.