

Why Did You Really Do It? Human Reasoning and Reasons for Action

José Ángel Gascón*

Received: 26 July 2020 / Revised: November 1 2020 / Accepted: 23 November 2020

Abstract: During the last decades several studies in cognitive psychology have shown that many of our actions do not depend on the reasons that we adduce afterwards, when we have to account for them. Our decisions seem to be often influenced by normatively or explanatorily irrelevant features of the environment of which we are not aware, and the reasons we offer for those decisions are a posteriori rationalisations. But exactly what reasons has the psychological research uncovered? In philosophy, a distinction has been commonly made between normative and motivating reasons: normative reasons make an action right, whereas motivating reasons explain our behaviour. Recently, Maria Alvarez has argued that, apart from normative (or justifying) reasons, we should further distinguish between motivating and explanatory reasons. We have, then, three kinds of reasons, and it is not clear which of them have been revealed as the real reasons for our actions by the psychological research. The answer we give to this question will have important implications both for the validity of our classifications of reasons and for our understanding of human action.

Keywords: Cognitive psychology; explanation; justification; motivation; rationalization; reasons for action.

* Universidad Católica del Maule

 <https://orcid.org/0000-0001-5571-6602>

 Department of Philosophy, Universidad Católica del Maule. Avenida San Miguel 3605, Talca, Chile.

 jgascon@ucm.cl

1. Introduction

Human beings consider, at least sometimes, what reasons we have to do something. When we do, according to a widespread view, what happens is the following: we usually act in the light of those reasons that seem to us to be the best, and then justify our action before others by putting forward the reasons that moved us to act. Imagine, for example, that I have been offered two jobs, one of which has a better salary and the other is in a city that I like. Which job I will accept depends on my weighing of those reasons (and others) and of which considerations are more important to me. After my decision, if I am challenged to justify it—“why did you do that?”—I will present those reasons that made me opt for one job rather than the other, hoping that my audience will see why I made the best choice. I will, then, engage in argumentation in order to show that those reasons that moved me to act in a certain way were the best reasons, all things considered.

This can be considered as the standard, common-sense view of reasoned action and justification of actions. We justify our actions by presenting reasons, and those are precisely the reasons for which we acted. Characterisations of the rational agent or the critical thinker which focus on reasons—as opposed to those which focus on the suitability of means to an end, for instance—tend to rely on this view of the relationship between reasons and action. According to Harvey Siegel, for example, a critical thinker is someone who is “appropriately moved by reasons” (1997, 49). And, in the literature on normativity and practical reasons, authors such as Scanlon (2014) and Kieseewetter (2017) define rationality in terms of responsiveness to reasons.

The idea that we should justify our actions by putting forward the reasons for which we acted seems like a plausible one. After all, there is a tendency to see people as irrational—or, at the very least, hypocritical—when they act for one reason and afterwards attempt to justify their action by appealing to different reasons. Consider the case of someone who decides to study philosophy and holds that her reason for that decision is her love of knowledge, when in fact what moved her towards a philosophical career is her desire to enjoy a high cultural status. No doubt many of us would see

that behaviour as falling short of rationality—or, if she is aware of her real motivation, as insincere.

However, if this is how we should understand the rational justification of actions, then apparently we are in serious trouble. The empirical research in the psychology of reasoning has shown that human beings are very bad at identifying the causes of *our own* actions. A growing number of empirical studies have provided evidence that we lack access to the knowledge of what considerations move us when we act. The reasons that we put forward when we are challenged to justify our behaviour are not, it seems, those reasons *for which* we acted, but merely our best guesses about why we acted that way—even though no doubt those guesses are sometimes right.

How worried should we be by this conclusion? The purpose of this article is to give a tentative answer to this question. I believe that any such answer, if it is to be plausible, must be both philosophically and psychologically informed. Our philosophical accounts of practical reasoning need to take into account the empirical findings that indicate what feats human reason can and cannot achieve; and, at the same time, the psychological research must be based on a philosophical understanding of reasons so that it is clear what conclusions can and cannot be drawn from the empirical data. I will begin, in the next section, by reviewing the empirical studies in psychology of reasoning that cast doubt on our ability to detect the reasons that move us to act. Then, in section 3, I will present philosophical distinctions between kinds of reasons and I will provide an interpretation of the conclusions of psychological studies in the light of those distinctions. Finally, in section 4, I will draw some preliminary conclusions about how all this should affect our conceptions of justification of actions and of rationalisation.

2. Psychological research on reasons for action

Up until the 1970s, it was widely assumed by psychological researchers that we are aware of the mental processes that lead to our judgements and our behaviour (Kunda 1999, 265). In order to study people's choices and evaluations, investigators resorted to self-report questionnaires in which the participants in the experiments were asked to state why they behaved as they did. Researchers who attempted to study the grounds for voting for a

political candidate or for choosing a job, for example, simply asked people why they voted for a certain candidate or why they chose a certain job. However, it eventually became manifest that such self-reports are not reliable.

In their ground-breaking article, Nisbett and Wilson (1977) reviewed a series of empirical studies in which a particular stimulus demonstrably influenced the participants' actions and judgements but, when interviewed, the participants denied that influence and tended to explain their behaviour by reference to other factors. An example is the large number of experiments that showed the existence of the "bystander effect," the fact that people are less likely to help a person in distress if there are many other onlookers around (Latané and Darley 1970). After the experiments, Latané and Darley asked the participants whether their decision to help or to abstain from helping had been influenced by the presence of other people. Despite the robust evidence that showed that a greater number of onlookers correlated with a failure to help, the participants systematically denied that influence. As the authors explain (Ibid., 124):

We asked this question every way we knew how: subtly, directly, tactfully, bluntly. Always we got the same answer. Subjects persistently claimed that their behavior was not influenced by the other people present.

Nisbett and Wilson also conducted a series of small studies in order to investigate the accuracy of causal explanations of one's own behaviour (Nisbett and Wilson 1977; Wilson and Nisbett 1978). The experiments were designed in a way that resembled as close as possible situations of the real life, with little or no deception involved. Yet they were also designed so that the stimuli that would probably influence the participants' behaviour were of a counter-intuitive sort and hence their influence could not be accounted for by the participants' prior causal theories of how people behave (Nisbett and Wilson 1977, 242). Therefore, those stimuli could only have been detected by the participants if they had genuine introspective access to their own cognitive processes. As expected, people were influenced by factors whose influence they could not detect—and, interestingly, the researchers themselves were highly unsuccessful in their predictions of which factors would influence them.

In one of those studies (Ibid., 243), the participants were asked to evaluate four pairs of stockings. They had to choose one of those pairs and, afterwards, they were asked why they had chosen it. The trick was that all the stockings were identical. Nisbett and Wilson observed that the stockings situated towards the right were preferred over the ones situated at the left. However, when the participants were asked about the reasons for their choices, the position of the article was never mentioned. In fact, when the researchers suggested that possibility to the participants, they denied it. The authors explain that (Wilson and Nisbett 1978, 124):

Only a quarter of the subjects required any prompting to explain the basis of their choices. Most of the subjects promptly responded that it was the knit, weave, sheerness, elasticity, or workmanship that they felt to be superior. [...] Not a single subject mentioned the position of the stockings as a reason for the choice.

Not only do we often fail to detect factors that cause our behaviours, but we also tend to report as reasons for our choices and judgements stimuli that actually had no effect on us. For example, in another experiment (Nisbett and Wilson 1977, 246), the participants had to predict how much electric shock they would take. Some of them were said that the shocks would do “no permanent damage,” while the others were not given that reassurance. Then, the researchers asked the first group whether that comment had affected their predictions, and they asked the second group whether, had they made that comment, their predictions would have been different. Inclusion of the reassurance proved to have no effect on the predictions of how much shock the participants would take, but a majority of them reported that it affected their predictions.

What all this evidence shows is not merely that we are sometimes wrong when we report our reasons for our decisions and judgements—that would hardly be big news. Neither can it be concluded that we are *always* wrong; as Nisbett and Ross (1980, 211) admit, we are often accurate in our explanations of the reasons for our behaviour. The worrying implication of that research on self-reports is rather that we *lack introspective access* to the reasons that guide our behaviour. The process by which we arrive at a belief of why we did something is the same whether that belief is accurate or inaccurate: we *infer* it from the known data and from our prior theories of

human behaviour. That is, it is the same process that we follow when we propose causal explanations of *other people's* behaviour (Ibid.). If, for example, I buy a bottle of water and I claim that I did so because I was thirsty, I am surely right. But this is so simply because we have a common-sense theory of why people usually buy bottles of water, and that theory is largely correct. Notice, also, that it would be just as easy to identify the reason why someone else bought a bottle of water. This was Nisbett and Ross's conclusion (1980, 211):

Empirically, this means that under most circumstances subjects will be right in their causal accounts if and only if observers, working with similar externally available information, also are right.

The problems begin when there is no prior theory or when that theory does not fit the case at hand. If we fail to help a person in distress because there are many other people around, or if we choose a pair of stockings because they are situated on the right, then we are likely to give a wrong account of our behaviour, since we have no prior theory about the relationship between those reasons and those actions. And, in those cases, we are just as likely to be wrong about our own behaviour as we are to be wrong about other people's behaviour. The process is the same in both cases.

Of course, when it comes to our own behaviour we have access to data that we lack when we attempt to interpret someone else's behaviour, such as our feelings, explicit goals, beliefs or memories (Nisbett and Ross 1980, 203). However, Wilson argues that this private information can also mislead us. He points out that "the vast amount of inside knowledge we have about ourselves increases confidence in our self-knowledge, but does not always lead to greater accuracy" (Wilson 2002, 113). A stranger with no access to that information could be more accurate about the causes of our actions, and in fact this seems to be often the case. He concludes (Ibid., 112):

Averaging across several studies, there seems to be no net advantage to having privileged information about ourselves: the amount of accuracy obtained by people about the causes of their responses is nearly identical with the amount of accuracy obtained by strangers.

If that conclusion is correct, then many of the processes that cause our actions and judgements are unconscious, just as the processes that are responsible for perception or textual comprehension. Kunda (1999, 270ff.) reviews other studies that provide evidence of those unconscious processes that influence our behaviour, including aspects such as implicit memory and subliminal perception. Some cognitive scientists have accepted the most dramatic implications of this conclusion regarding our conscious will. Evans, while admitting that there is a difference between voluntary and involuntary actions, questions the very existence of a conscious will (2010, 177):

‘We’ are not conscious persons in control of our behaviour and the reflective mind does not equal a conscious mind. The conscious person is a construction of the brain, an illusory narrative that accompanies us through life.

In the same vein, Wegner (2002) talks about the conscious will as an “illusion.” According to his theory of apparent mental causation, conscious will is not a cause of actions but simply a (possibly misguided) *feeling* that an action was caused by us. He explains (Ibid., 336):

Apparent mental causation suggests that the experience of consciously willing an act is merely a humble estimate of the causal efficacy of the person’s thoughts in producing the action. Conscious will is the mind’s way of signaling that it might have been involved in causing the action. The person’s experience of doing the act is only one source of evidence regarding the actual force of the person’s will in causing the action, however, and it may not even be the best source.

Although Wilson does not endorse the conclusion that conscious will is *always* an illusion, he admits that very often it is (2002, 48): “We may have the impression that we, our conscious selves, are in complete control, but that is at least in part an illusion.”

Now, if we accept these psychologists’ conclusion that we tend not to be (or perhaps never are) introspectively aware of what factors influence our actions and judgements, and in fact we are often wrong about them, the question is: how big a problem is that for our philosophical theories about

reasons for action and justification of actions? This is a very broad issue that cannot be solved in a single paper. As a first step, however, it would help to be clear about what exactly Nisbett and Wilson's experiments uncovered. Did they identify our *real* reasons for action? Or did they show us simply the *causes* of our actions? Are they the same thing? Sorting out this conceptual issue is the purpose of this paper, and to this I move in the next section.

3. What reasons are we talking about?

The results of the experiments conducted by Nisbett and Wilson certainly seem to reveal something important that jeopardises our ideas of intentional action and justification of actions. But, what is it exactly that was identified in those experiments? In their articles, Nisbett and Wilson used a variety of terms to refer to the stimuli that influenced the participants' behaviour: "influences," "explanation," "causes," "causal factors," and "reasons" for choice. The point was that there seemed to be a mismatch between the reasons stated by the participants in the studies and whatever it was—influences, reasons, causes—that explained their choices. Thus, a necessary first step in the assessment of the implications of those studies for our philosophical theories is the clarification of these factors that explained the participants' behaviour.

The most natural interpretation, I believe, is that the experiments identified the *causes* of our actions and judgements. Now, it is well known that, according to some philosophical views, reasons for action just are the causes of our actions. Davidson (1963) famously argued for that view. If that is how we should understand practical reasons, then the discovery that people lack direct awareness to the causes of their actions obviously challenges our practice of giving reasons for our actions. If we cannot detect the causes of our behaviour, and practical reasons are precisely those causes, then it seems that the reasons we give for our actions are mere speculations. In that case, we cannot be sure for what reasons we did something, just as we cannot be sure how our stomach is digesting what we ate.

However, the philosophical literature has distinguished between different kinds of reasons, and Davidson focused on only one of them: the kind

of reason that “explains the action by giving the agent’s reason for doing what he did” (Ibid., 685). Beyond reasons that explain the motivations of agents, there are also reasons that justify their actions. It is one thing to report what considerations *motivated* us to do something, and hence explain our action; it is something different to *justify* our actions with considerations that make them the right thing to do (Dancy 2000, 20–25). The former kind of reasons has been called *motivating* reasons, whereas the latter has been called *normative* reasons. Thus, Parfit (1997, 99) says that normative reasons are those that we are looking for when we ask “What do we have most reason to want, and do?”; motivating reasons, on the other hand, are those in light of which we act. Dancy explains the distinction this way (2000, 2):

There is the question what were the considerations in the light of which, or despite which, he acted as he did. This issue about *his reasons for doing it* is a matter of motivation. There is also the question whether there was good reason to act in that way, as we say, *any reason for doing it* at all, one perhaps that made it sensible in the circumstances, morally required, or in some other way to be recommended, or whether there was more reason not to do it. [...] This second question raises a normative issue.

We can act for a good reason, in which case our motivating reason is also our normative reason, but it is also possible for these two kinds of reasons to diverge. Imagine, for example, that I voted for a certain political candidate because it seemed to me that she was the most honest and competent one. Those were the reasons that I considered when I was deciding my vote, so those are the reasons for which I acted. When asked, I offer those reasons to justify my choice. In this case, my normative reasons are the same as my motivating reasons. But let us imagine a slightly different scenario. Imagine that, even though that political candidate was indeed the most honest and competent one, I did not take that fact into account when deciding my vote; instead, what motivated me to vote for her was that she was born in the same city as me. I still justify my vote before others by mentioning her honesty and competence, but I know that I voted for her because we were born in the same place. In this second case, my motivating reasons are different from my normative reasons.

Let us differentiate, then, between:

Normative reasons: Considerations that make an action the right thing to do, that count in favour of doing that action.

Motivating reasons: Considerations that moved me to do something, those in the light of which I acted.

As Dancy (2000, 2) and Alvarez (2009) argue, this reference to two “kinds” of reasons should not be understood as implying that there are really two sorts of reasons—reasons that motivate and reasons that justify. They are different kinds of reasons only in the sense that they are offered in answer to two different questions: (1) what makes that action right?, and (2) why did you do that action?

Now, if we go back to Nisbett and Wilson’s experiments and ask what kind of reasons—if any—they have discovered, it seems clear that we can rule out normative reasons. The researchers, as we have seen, deliberately designed the experiments so that there were no good reasons to prefer one pair of stockings rather than another—they were all of the same quality. The reasons that the participants gave—the superior knit, weave, sheerness, elasticity, or workmanship—were false, and it is a commonplace in philosophy that bad normative reasons are not reasons at all¹. If anything, Nisbett and Wilson showed that the participants in the experiments could not offer any normative reasons. The experiments certainly did not uncover any normative reasons for there was none in those cases.

Yet, it is not obvious to me either that the findings of the experiments refer to motivating reasons. Those findings do refer to factors that explain people’s actions, but motivating reasons are not simply any kind of explanation; motivating reasons explain actions only insofar as those actions were made *in the light of* reasons. That means that a causal factor would not count as a motivating reason if the agent has not consciously considered it and decided to act on the basis of it. For a cause of people’s actions to be

¹ Dancy puts it at the very beginning of his book (2000, 1-2): “A bad reason for doing something, if it is not merely a not very good reason for doing it, can only be no reason at all for doing it; if so, it is not a reason in the sense intended, since it does not favour the relevant action.”

a motivating reason, they must at least recognise it as a reason and be guided by it. Suppose, for instance, that I fell on the street because a car hit me. It does not make sense to say that my motivating reason for falling was that a car hit me, i.e. that I was motivated to fall by the hit of a car. No doubt the hitting of the car explains the event, but it is not an explanation in terms of *motivation*.

There is, however, a third possibility besides normative and motivating reasons: what Searle (2001, 111) calls “straight causal explanations” and Dancy (2000, 5) calls “reasons why.” These reasons do not involve considerations that the agent takes to favour some action. Actions that are explained by “reasons why” are not performed in light of those reasons, but simply caused by them. This is the case with the explanation of the fact that I fell on the street that mentions the hit of the car. Many other events involve this kind of explanations, in which no reason was considered by the agent, as Dancy reminds us (*Ibid.*):

What explains why one person yawned may be that someone else yawned just next to them. What explains why he responded so aggressively may be that he is having trouble at home or that he has taken a particular form of medication. What explains why he gave this student a better grade than she deserved is that he was unconsciously influenced by the fact that she always dresses so neatly (or something even less defensible). What explains why so many people buy expensive perfume at Christmas is the barrage of advertising on the television. What explains why he didn't come to the party is that he is shy. In none of these cases are we specifying considerations in the light of which these things were done.

Dancy states that what these explanations involve “is not a reason at all, really, but rather a cause” (*Ibid.*, 6). However, Alvarez (2009, 184) argues that its being a cause does not exclude its being a reason, since both terms belong to different domains: that of causation and that of explanation. We use reasons to explain actions, and those reasons sometimes happen to be causes in the natural realm. Therefore, she proposes that, besides normative and motivating reasons, we should consider *explanatory* reasons. If we differentiate among different kinds of reasons on the basis of the role

they play in answering different questions, then the question of what explains an action is substantially different from the question of what motivated the agent—even though, of course, the same reason can answer both questions.

We have, then, according to Alvarez’s proposal, three kinds of reasons:

Normative reasons: Considerations that make an action the right thing to do, that count in favour of doing that action.

Motivating reasons: Considerations that moved me to do something, those in the light of which I acted.

Explanatory reasons: Considerations that explain why I did something, what caused my action.

Explanatory reasons are a better candidate for the kind of reasons that the psychological experiments revealed. They are causes that explain the participants’ actions without being at the same time motivating reasons, since the participants did not consider them and even denied their influence. Those reasons are causes in the same sense that taking a certain medication is the cause of aggressive behaviour. They influence our actions but we are unaware of that influence.

There is one crucial difference between explanatory reasons and the other two kinds of reasons, and that difference is what makes the findings of psychological experiments so shocking: explanatory reasons do not necessarily involve *human agency*. Just as they can be used to explain human actions, they are also what explains events such as the rain, the collapse of a building or the movement of waves at sea. There are no normative or motivating reasons for events like these—water and buildings do not consider reasons and do not attempt to justify their actions. So, when human actions are explained on the basis of explanatory reasons that are not also normative or motivating, that certainly feels like our sense of agency itself is being challenged. That may be all right for certain human actions, such as yawns or sudden outbursts of aggressiveness, but it is frightening to find out that it also involves actions for which we believe we have motivating reasons, such as choosing stockings of helping a person in distress. No wonder some cognitive scientists have concluded that conscious will is an illusion.

Should we then give up any talk of reasons altogether? Even though satisfactorily solving this issue would require a longer discussion, and the main topic of the present article was to sort out the kinds of reasons that the psychological experiments are referring to, in the next section I will outline a path that—in my view—we should take. In order to sketch an answer that question, I believe we must go beyond Nisbett and Wilson’s experiments and consider the role of a kind of reasons that initially did not seem empirically relevant: normative reasons.

4. Justification, motivation and rationalisation

Normative reasons are importantly different from explanatory and motivating reasons. The question of what considerations count in favour of an action, what considerations make an action right, is not empirical but normative. The psychological research can test our conceptions of explanatory and motivating reasons—of what explains our actions and what motivates us—but only the philosophical reflection can test our views on normative reasons. What is right is right regardless of whether it explains our actions or motivates us. That is why Nisbett and Wilson’s experiments could not throw any light on normative reasons.

However, what interests us here is not merely whether normative reasons exist. For normative reasons to be something more than a philosophical construct, they must influence our actions somehow. If the reasons which justify our actions have no influence on our behaviour, as the experiments that we have seen suggest, then it begins to look as if those reasons were merely *epiphenomenal*: they would play no role in the determination of our actions.

How could we measure the causal efficacy of normative reasons? The safest way, I believe, is to focus on the reasons that we offer with the deliberate purpose of justifying actions. Even though we often attempt to justify our actions by explaining why we performed them—i.e. by citing motivating reasons—normative reasons need not be also motivating reasons (Dancy 2000, 3). Sometimes we simply argue that what we did was right without intending to explain what moved us to do it. Someone might, for example, argue that the choice of her academic career was a good one—because, say,

it has good job prospects and it fits her character—without even remembering why she chose it in the first place. When we attempt to justify an action this way, what we offer is *purported* normative reasons—these include both genuine normative reasons and bad reasons, which, as was pointed out in the previous section, cannot be considered reasons. Sometimes the reasons with which people attempt to justify their actions are good ones, and sometimes they are false and hence they are not really reasons. Normative reasons, therefore, are a *subset* of the purported normative reasons that people can offer.

Thus, my main claim in this section is the following: if the reasons that we consider and offer to justify actions play a causal role in our behaviour, then it follows that normative reasons are causally efficacious. That is, if purported normative reasons influence our behaviour, and normative reasons are a subset of purported normative reasons, then normative reasons must influence our behaviour. In plain words, if our actions are influenced by reasons which we think (correctly or incorrectly) that would justify our actions, then it can be said that at least sometimes our actions are influenced by reasons that do justify our actions. It would be very odd indeed if we were influenced by reasons but only the bad ones.

In fact, we have evidence that shows that at least sometimes purported normative reasons motivate our decisions and beliefs. The idea that people can take decisions and change their minds on the basis of reasons that show that some action is the right thing to do seems to be a necessary assumption in order to account for much of human behaviour. This can be seen most clearly in psychological experiments involving interpersonal argumentation. As Mercier and Sperber (2017, 264–265) point out, groups of people are more able to solve logical problems than individuals working alone, and this happens because people working in groups benefit from the exchange of reasons. For example, Trouche, Sander and Mercier (2014) showed that people who are confronted with arguments or who argue are more likely to solve logical problems such as those of the Cognitive Reflection Test (Frederick 2005) and others. Their experiments were designed in a way that ruled out the effect of degrees of confidence of some participants on others, measuring specifically the effects of good argumentation. Thus, they concluded that their results “make it clear that arguments, rather than confidence, are

the main factor explaining the performance of groups discussing intellectual tasks” (Ibid., 1968).

According to Mercier and Sperber, the main function of the human faculty of reason is not to make better decisions but—at least sometimes—to make decisions for which we can come up with (allegedly) good reasons (Mercier and Sperber 2017, 255): “when people have weak or conflicting intuitions, reason drives them toward the decision for which it is easiest to find reasons—the decisions that they can best justify.” According to their argumentative theory of reasoning, the justifications that we offer or that we mentally rehearse do guide our actions. If purported normative reasons are understood as attempts to justify actions—as I have assumed here—then they seem to influence decisions. Purported normative reasons can influence us in group discussion, as Trouche et al. showed; or, even when there is no interpersonal argumentation taking place, the prospective justification that we mentally rehearse leads us in the direction of the most acceptable reasons.

It may be thought that this conclusion clashes with certain experiments that show that rational argumentation rarely changes people’s minds, particularly in the moral realm (Haidt 2001). I admit that sometimes that may be the case and that the power of normative reasons is somehow limited. Nevertheless, this does not mean that they never have any effect. Cohen (2019, 715) addresses this problem and notes that:

The point is that the marginalization of reason as only *rarely* effective is also an acknowledgement that it sometimes is effective. The claim is not that there is no causal footprint for reasoning and argumentation at all; rather, the claim is that the effects are limited.

In fact, just as there is evidence that arguments often fail to convince people in certain domains, we have also evidence that sometimes—not infrequently, I would say—arguments change people’s minds. As we have seen, the experiments that Mercier and his collaborators carried out and reviewed show exactly that. Moreover, I would like to add the observation that, even when the participants in an experiment fail to be convinced by arguments when they should, all the experiments have shown is that people have not been convinced *immediately*. There is still the possibility that

people keep thinking about the reasons they have heard and change their minds *in the long run*. In fact, Kjeldsen (2020) interviewed people about their experiences of changing their minds on important social or political issues and he found out that it took them between 4 and 9 years to do so.

Hence, normative reasons do not seem to be inert. They can lead people to take a decision or form a judgement, even if it takes time. But then, when such a thing happens, we can confidently say that those are people's *motivating* reasons. Just as in Nisbett and Wilson's experiments there was no way that the motivating reasons reported by the participants played any role in determining behaviour, in the experiments reviewed by Mercier and Sperber there seems to be no alternative to granting normative reasons a causal role. Therefore, reasons are not a mere epiphenomenon; motivating reasons, i.e. conscious reasons with a causal power on our decisions, exist.

Concluding that motivating reasons exist, however, is not much if those reasons can only be reliably detected in the laboratory. And here lies precisely the lesson that we should draw from the psychological research: we do not have direct access to the causes of our own actions, we just infer the possible causes from a body of data and a more or less accurate theory of human behaviour, so we can always be wrong about our alleged motivating reasons. We should not be confident that we did something for the reasons that we think we did it. We need to accept our unreliability even in the realm of our own actions. As Wilson suggests (2002, 113): "we all might want to be more humble about the accuracy of our causal judgments."

So one lesson regarding our own self-reports is that we should acknowledge the possibility that *we* are wrong. What about other people's accounts? In my view, taking that conclusion seriously should lead us to giving considerably less weight to motivating reasons in people's attempts to justify their actions. Accounts of why people did something should not be given a predominant place in justifications of actions. When it comes to justification, we should focus on normative reasons, and these should be kept separate from motivating reasons. I believe this is a conclusion that we should accept in light of the unreliability of our reports of motivating reasons. If we do not want the weakness of those reports to be transferred to our practice of justifying actions, the kind of reasons that make an action

right or wrong should be relatively independent of the kind of reasons that explain why that action was done.

This may seem too radical a proposal, as it risks blurring the distinction between a genuine justification of an action and a *rationalisation*. Without that distinction, it may be thought that the very idea of rationality is in danger. I will use the rest of this section to try to dispel that worry.

In the most common sense of the term, a rationalisation is a purported account offered by an agent of one of her actions that (Audi 1985, 163):

Offers one or more reasons for doing that action.

Represents his doing that action as at least *prima facie* rational given those reasons.

Does not explain why the agent did that action.

That is, a rationalisation is an attempt to *justify* an action (point 2) by offering *normative* reasons (point 1) that are not at the same time *motivating or explanatory* reasons (point 3). Someone might, for example, justify his decision not to eat peppers by asserting that they are bad for his health, when in fact his motivating reason is simply that he does not like them. But rationalisation so defined is exactly what, I have argued, theories of rational action should be more tolerant of. Should we then accept the fact that, as Mercier and Sperber (2017, 253) say, humans are rationalisation machines? In that case, it seems that we would be condoning widespread irrationality. The problem with the empirical studies that show that rationalising is what the human mind usually does is that they seem to warrant the conclusion that we are all irrational. As Cohen (Cohen 2019, 711) puts it:

[...] a great deal, perhaps even most, of our reasoning turns out to be *rationalizing*. The reasons we give for our positions are seldom either the real motives or the effective causes of why we have those positions. The uncomfortable conclusion, unfortunately substantiated by too many empirical studies to dismiss, is that we are not as rational as we like to think.

However, I believe that we can dissipate (at least most of) the fear of irrationality if we do not underestimate the extent to which normative

reasons can be criticised. Justifications, even if they are rationalisations, can still be correct or incorrect. A match between purported normative reasons and motivating reasons is not the only way to check the correctness of justifications—it is not, in fact, the main one or even the most demanding one. Purported normative reasons by themselves must fulfil several criteria for them to constitute a satisfactory justification. One of those criteria is, of course, that they must be true, they must mention real facts. This criterion alone allows us to see where the participants in Nisbett and Wilson’s stockings experiment got their justification wrong: they mentioned particular features that allegedly made certain stockings the best ones in the lot, whereas in fact all of them were identical. There is no need to appeal to their motivating reasons in order to conclude that their justifications were flawed.

Besides truth, we should also expect an agent’s purported normative reasons to *cohere* with those that the same agent has offered in similar circumstances. The principle that like cases should be treated alike is firmly established both in law and in ethics, but it is also relevant in other domains. This principle helps us explain what might be wrong in the justifications offered by the participants in the bystander effect experiments performed by Latané and Darley. Surely, those who helped the person in distress when there were few onlookers could have justified their action by saying that the person needed help, but if they do not help in a similar scenario with more onlookers, there might be an incoherence in the normative reasons they state.² Again, as in Nisbett and Wilson’s experiments, we can talk about the rationality or irrationality of justifications without checking motivating reasons.

Consider, finally, a common example that Audi mentions (1985, 159–160): “a person cites an altruistic reason he had for helping someone, when in fact he was motivated by selfish reasons.” If what explains that action is selfish reasons, one would expect that the person would not behave the

² I say that there *might* be an incoherence because I am not sure that there is no relevant difference between the two scenarios to which the agent could rightly point out. After all, if there are many onlookers, the agent could always argue that she thought that someone would take care of the person in distress, and perhaps that could be a legitimate expectation.

same way in a situation in which she again must help someone but the selfish reasons are absent—there is no benefit for her. That would reveal an incoherence in her attitude towards purported normative reasons between both cases. Otherwise, if her behaviour is *consistently* helpful, it seems to me that insisting on the existence of a selfish motivation in order to criticise her reasons would entail a moral theory that is way too demanding.

All this is not intended to mean that we should *never* take into account motivating reasons when assessing justifications. Even within the boundaries of a single action, *sometimes* a manifest mismatch between purported normative and motivating reasons can be reprehensible. If it is clearly apparent, for example, that I intended to punch someone out of anger and, by sheer luck, I ended up moving him away from a bus that was going to run over him, thus saving his life, then I can hardly justify my action by saying that I saved his life. Anyone could see that my intention was to hit him. However, apart from clear cases like this one, our practice of giving and asking for justifications should not focus on mismatches between purported normative, motivating and explanatory reasons. We should accept that those mismatches are ubiquitous in human action, as the research in experimental psychology has shown, but at the same time we can be confident that we have the resources to assess justifications by themselves.

5. Conclusion

Research in cognitive psychology during the last five decades has shown that, in many situations, the reasons with which people explain their own actions and judgements do not correspond to the real factors that caused them. This finding has led to the conclusion that people do not have introspective access to the causes of their own behaviour; instead, people infer them, just as they would if they were observers of someone else's behaviour. Such a conclusion seems to cast doubt on the significance of our practice of justification of actions and exchange of reasons. However, in order to fully understand the philosophical implications of the results of psychological research, we need to be clear about what kinds of reasons psychologists are talking about.

In the philosophical literature, three kinds of reasons have been distinguished, according to the kind of question that they answer: *normative* reasons, considerations that make an action right; *motivating* reasons, considerations in light of which the person acted; and *explanatory* reasons, considerations that explain what caused an action or an event. The problem with the psychological experiments, we saw, was that the participants offered purported motivating reasons that did not explain their choices at all; instead, what explained their choices were explanatory reasons that the experiments uncovered and of which the participants were unaware. That challenges the reality of motivating reasons, and we are left only with normative reasons that, for all we know, could have no effect on behaviour whatsoever—they could be epiphenomenal.

However, we also saw that certain behaviours could only be plausibly accounted for by the influence of purported normative reasons. If our performance in a logical task is better when there is argumentation, and if we tend to lean towards the most justifiable decisions when our intuitions about what to do are weak, that gives us grounds for believing that sometimes purported normative reasons do guide our actions. If that is the case, then actual normative reasons—being a subset of purported normative reasons—must at least sometimes influence our behaviour. The problem, given our lack of introspective access to the causes of our behaviour, is that in practice we can never be sure that, in a particular instance, we are genuinely motivated by normative reasons. For this reason, I argued that our assessments of the reasons produced by agents should not give much weight to whether they are also motivating reasons or not—i.e. whether they are rationalisations of actions. Outside laboratory conditions, the identification of motivating reasons is a tricky issue and it is bound to lead to speculations, and we have the conceptual resources to assess purported normative reasons in themselves.

Acknowledgements

A previous version of this paper was presented at the 12th Conference of the Ontario Society for the Study of Argumentation (OSSA), which took place on-line on 3-6 of June of 2020. Special thanks to my commentator, Marcin

Koszowy, for his useful observations. I should also thank the people who attended my talk and offered insightful suggestions, especially Daniel Cohen, Hans Hansen and Júlder Gómez.

Funding

This research was possible thanks to the postdoctoral scholarship FOND-ECYT 3190149 of ANID/CONICYT, and also to the project PGC2018-095941B-I00, “Prácticas argumentativas y pragmática de las razones,” of the Spanish Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación.

References

- Alvarez, Maria. 2009. “How Many Kinds of Reasons?” *Philosophical Explorations* 12 (2): 181–93. <https://doi.org/10.1080/13869790902838514>
- Audi, Robert. 1985. “Rationalization and Rationality.” *Synthese* 65 (2): 159–84. <https://doi.org/10.1007/BF00869298>
- Cohen, Daniel H. 2019. “Argumentative Virtues as Conduits for Reason’s Causal Efficacy: Why the Practice of Giving Reasons Requires That We Practice Hearing Reasons.” *Topoi* 38 (4): 711–18. <https://doi.org/10.1007/s11245-015-9364-x>
- Dancy, Jonathan. 2000. *Practical Reality*. New York: Oxford University Press.
- Davidson, Donald. 1963. “Actions, Reasons, and Causes.” *Journal of Philosophy* 60 (23): 685–700. <https://doi.org/10.2307/2023177>
- Evans, Jonathan St. B. T. 2010. *Thinking Twice: Two Minds in One Brain*. New York: Oxford University Press.
- Frederick, Shane. 2005. “Cognitive Reflection and Decision Making.” *Journal of Economic Perspectives* 19 (4): 25–42. <https://doi.org/10.1257/089533005775196732>
- Haidt, Jonathan. 2001. “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review* 108 (4): 814–34. <https://doi.org/10.1037/0033-295x.108.4.814>
- Kiesewetter, Benjamin. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.
- Kjeldsen, Jens E. 2020. “What Makes Us Change Our Minds in Our Everyday Life? Working through Evidence and Persuasion, Events and Experiences.” In *OSSA Conference Archive*, 1–14. <https://scholar.uwindsor.ca/ossaarchive/OSSA12/Saturday/7>.

- Kunda, Ziva. 1999. *Social Cognition: Making Sense of People*. Cambridge, MA: MIT Press.
- Latané, Bibb, and John M. Darley. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century Crofts.
- Mercier, Hugo, and Dan Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press. <https://doi.org/10.4159/9780674977860>
- Nisbett, Richard E., and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, NJ: Prentice-Hall.
- Nisbett, Richard E., and Timothy DeCamp Wilson. 1977. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84 (3): 231–59. <https://doi.org/10.1037/0033-295X.84.3.231>
- Parfit, Derek. 1997. "Reasons and Motivation." *Proceedings of the Aristotelian Society, Supplementary Volumes* 71: 99–130. <https://doi.org/10.1111/1467-8349.00021>
- Scanlon, Thomas M. 2014. *Being Realistic about Reasons*. New York: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199678488.001.0001>
- Searle, John R. 2001. *Rationality in Action*. Cambridge, MA: MIT Press.
- Siegel, Harvey. 1997. *Rationality Redeemed? Further Dialogues on an Educational Ideal*. New York: Routledge.
- Trouche, Emmanuel, Emmanuel Sander, and Hugo Mercier. 2014. "Arguments, more than Confidence, Explain the Good Performance of Reasoning Groups." *Journal of Experimental Psychology: General* 143 (5): 1958–71. <https://doi.org/10.1037/a0037099>
- Wegner, Daniel. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/3650.001.0001>
- Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.
- Wilson, Timothy de Camp, and Richard E. Nisbett. 1978. "The Accuracy of Verbal Reports about the Effects of Stimuli on Evaluations and Behavior." *Social Psychology* 41 (2): 118–31. <https://doi.org/10.2307/3033572>