Organon F 25 (1) 2018: 53-73

# Radical Rationalization Accommodates Rampant Irrationality

JOACHIM LIPSKI[1]

ABSTRACT: According to a classic position in analytic philosophy of mind, we must interpret agents as largely rational in order to be able to attribute intentional mental states to them. However, adopting this position requires clarifying in what way and by which criteria agents can still be irrational. In this paper I will offer one such criterion. More specifically, I argue that the kind of rationality methodologically required by intentional interpretation is to be specified in terms of psychological efficacy. Thereby, this notion can be distinguished from a more commonly used notion of rationality and hence cannot be shown to be undermined by the potential prevalence of a corresponding kind of irrationality.

KEYWORDS: Cognitive biases – holism – intentionality – interpretation – rationality – reason-explanation.

## 1. Introduction

Some fields, with intentional psychology and economy chiefly among them, methodologically require the assumption that human beings are "rational animals". At the same time, claims that irrationality runs rampant in

[1] Received: 24 March 2017 / Accepted: 2 August 2017

✉ Joachim Lipski

Research Center for Neurophilosophy and Ethics of Neuroscience
Ludwig-Maximilians-University Munich
Schellingstr. 10/III, Room J 308, 80799 Munich, Germany

e-mail: joachim.lipski@gmx.de

the general population are not hard to find. For example, Bertrand Russell lamented:

> Man is a rational animal – so at least I have been told. Throughout a long life, I have looked diligently for evidence in favour of this statement, but so far I have not had the good fortune to come across it, though I have searched in many countries spread over three continents. (Russell 1950, 71)

In a more timely treatment of the topic in current cognitive science, Stanovich echoes this sentiment when he remarks that due to irrationality,

> [P]hysicians choose less effective medical treatments; people fail to accurately assess risks in their environment; information is misused in legal proceedings; millions of dollars are spent on unneeded projects by government and private industry; parents fail to vaccinate their children; unnecessary surgery is performed; animals are hunted to extinction; billions of dollars are wasted on quack medical remedies; and costly financial misjudgments are made. (Stanovich 2003, 293; also cf. Stanovich 2009, 197 f.)

In this paper, I will explore whether said potentially rampant irrationality can conflict with the intentional (or "folk-psychological") method of mental explanation. Prima facie, the assumption that there is such a conflict is motivated by a classic position in analytic philosophy, which holds that, in order to be able to explain a person's intentional mental states and actions and to be able to attribute the former and responsibility for the latter to her, we need to interpret her as by and large rational. In this paper, I will argue that mental state attribution by way of intentional interpretation and the position we might call rationality-skepticism (as just expressed in Russell's and Stanovich's quotes) track two distinct notions of rationality, and I will offer a criterion for distinguishing them. Specifically, the kind of rationality methodologically required by intentional explanation is tied to psychological efficacy relative to an agent's mindset, whereas the other is not.

To be sure, the possibility of irrationality is a classic and often-discussed topic. What I wish to add to the debate, which for present purposes

we can think of as having started with Davidson's famous entry (see Davidson 1980, 21-42; see my later fn. 10), is an explicit reconciliation of methodological rationalization and a common-sense theory of irrationality.[2] What we can find to dominate the literature are positions which take sides in this conflict by trying to demolish one of the opponents, either by denying the reality of irrationality (e.g. Bratman 1979; Buss 1997; Arpaly 2000) or by denying that explanation of minds require rationalization in the relevant sense (a denial of rationality; e.g. Mele 1987, 37). However, I believe there is room for acknowledging both; hence, I will argue for their reconciliation and the dissolution of their purported conflict. If anything, my strategy is similar to Holton's (1999), insofar as I think of irrationality as not touching methodological rationalization – but that is where the similarities between Holton's and my account already end.

## 2. Intentional explanation and rationalization

One way of explaining an agent's behaviour is by attributing intentional mental states to her (cf. Cummins 2000, 127 ff.; Fodor 1989, chap. 1). Following Davidson, a methodological prerequisite for intentional explanation is to construe the agent by and large – and as far as possible – as rational:

> [I]f we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behaviour, then we are committed to finding, in the pattern of behaviour, belief and desire, a large degree of rationality and consistency. (Davidson 1980, 237)[3]

In order to understand an agent, "we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own

---

[2] In recent discussions of cognitive biases, the term "rationalization" has been popularly used as referring to the act of trying to make one's actions appear rational after the fact (so-called "post-hoc rationalization"; cf. Sie & Wouters 2010). However, I only use "rationalization" in its Davidsonian sense, which simply means the act of interpreting agents as by and large rational. According to Davidson, it is methodologically required when attributing mental states (see my section 2).

[3] Regarding this point, see also Davidson (1980, 221 f.); Davidson (2001a, 196-200); Lewis (1983, 113); and Dennett (1987, 19, fn. 1).

lights, it goes without saying)" (Davidson 1980, 222). According to this view, we must attempt to assign beliefs and desires which, by rational standards, jointly produce the actions we can observe an agent to carry out. To say that they jointly produce the action by rational standards is to say that, when taking means-ends beliefs and desires as premises (or "primary reasons"; Davidson 1980, 3-18) in a practical syllogism (Broadie 1986), the conclusion, whose content is the respective action, is logically derivable from them.

According to Davidson (whose writings, along with Quine's, form the *loci classici* for methodological assumptions of rationality), the reason for why rationalisation is necessary is that in interpreting an agent we need to untangle her observable behaviour – which tracks what the agent holds true – into its two aspects, namely belief and meaning (Davidson 2001a, 148, 195). The only way to untangle this vector is by maximizing the truth of an agent's beliefs (or their *correspondence* with actual facts; see Davidson 2001a, 196) on the one hand and the *coherence* of her intentional states on the other (see Davidson 1980, 237). That is, in order to be ascribable to an agent, intentional states have to stand in rational relations to her other mental states, to her actions, and to the world. Hence, irrationality is limited:

> Coherence here includes the idea of rationality both in the sense that the action to be explained must be reasonable in the light of the assigned desires and beliefs, but also in the sense that the assigned desires and beliefs must fit with one another. The methodological assumption of rationality does not make it impossible to attribute irrational thoughts and actions to an agent, but it does impose a burden on such attributions. We weaken the intelligibility of attributions of thoughts of any kind to the extent that we fail to uncover a consistent pattern of beliefs, and, finally, of actions, for it is only against a background of such a pattern that we can identify thoughts (Davidson 2001a, 159).

As pointed out by Searle (2000, 106), intentional explanations work the way they do because their explananda are both rationally and causally derivable from the ascribed mental states: That is, if someone just drank a glass of water, then knowing that she was thirsty and that she believed she could quench her thirst by drinking a glass of water explains her drinking.

Invoking causal relations is necessary insofar as we cannot merely rely on rational relations to intentionally explain actions: For instance, the logical relations inherent in the syllogism "Drinking a glass of water is a way to quench thirst; she is thirsty; thus it would be reasonable for her to drink a glass of water" do not, by themselves, establish the required causal-psychological relation; i.e., the premise-conclusion-relations do not by themselves establish that the conclusion expresses a psychological motivation caused by the truth of the first and the instantiation of the second premise. Because even if it was in this sense rational for someone to drink a glass of water, two things may keep her from drinking: further reasons speaking against drinking or any type of external (i.e. non-mental) obstacle. We distinguish these two cases by saying that we either *decided* against drinking or that we were *kept* from drinking. The first implies rational control of the agent, the other a non-mental obstacle (which may be a brick wall just as much as a mental disease – meaning this obstacle does not have to be external to the body, but rather beyond agential control). It is the first case with which we are presently concerned, since it says something about the agent's mindset in regard to rationality whereas the latter does not. That is, even if a thirsty person eventually decides against drinking, the fact that she had a reason for drinking is not to be disregarded in an account of her mental state. Rather, we say that the agent had conflicting reasons, and if we wish to continue thinking of her as rational, we should say that the stronger reason won out and caused her not to drink.[4] This way, the form of a causal and rational explanation is maintained, even if some reasons (such as her thirst) ultimately proved not to be causally effective for her action.[5]

Reflecting these two components, causality and logicality, this kind of explanation is both nomological and normative: Its explanatory form, the practical syllogism, is formally analogous to deductive-nomological (DN)

---

[4] Sometimes, weaker reasons may win out, in which case we speak of akrasia or weakness of the will (cf. Davidson 1980, 21-42). I will not explore the specific issue of akrasia here.

[5] I tacitly join Searle (1979, 85-87) in assuming that reasons and causes need not be mutually exclusive. Further, I assume that intentional explanation is a quasi-scientific endeavour (cf. Sellars 1997, 90-117; Fodor 1989, 7; Davidson 1980, 221; Dennett 1991, 28 f.).

explanation (see Hempel & Oppenheim 1948), which also combines causality and logicality. For this reason, and for ease of exposition, I will henceforth call explanatory intentional psychological generalisations "psychological laws".[6] One reason for why psychological laws have explanatory value is that they can be found to be widely applicable. This means, firstly, that their explanatory categories have been and continue to be instantiated across many different individuals and events and, secondly, that their instantiations are generalizable (i.e. that they can be viewed *as instantiations* of a general law) due to commonly known, if usually implicit, criteria for what counts as having the kind of psychological attitude that is attributed (e.g. it is generally known that someone's raising their voice can provide evidence for attributing anger to them) (Cummins 2000, 127). Which is to say that reasons are psychologically efficacious: People act for reasons, their actions can be influenced by providing them with reasons, anyone's actions can be predicted on the basis of knowing her reasons for acting, and so on (for a defence of this view see Fodor 1989, chap. 1). Additionally, their being rational generalisations means that the explanatory value of psychological laws also depends on their gaining their general applicability not just by expressing causal relations, but by expressing rational ones: by stating what is rational, and by thereby either being normatively binding themselves or by descriptively referencing such norms of rationality.

A caveat is in order. Clearly, there are psychological laws stating general effects which are not rational, such as in the case of active implicit biases (see Tversky & Kahnemann 1974; Sunstein 2005; Gigerenzer 2008; Sie & Wouters 2010, 126-128.).[7] That the causes of irrational cognition

---

[6] By which I neither mean to imply that psychological explanation is as strict, or as general as other kinds of explanation usually associated with the DN account, nor that we can generally conflate the two notions. (For whether and in what way explanation in psychology is "lawlike" in a stronger sense see Cummins 2000 and Bechtel & Wright 2009.)

[7] It should be noted that there is a tendency in recent moral psychology to operationalize reason as a process of conscious deliberation which should immediately precede decisions, judgments or, generally, actions. However, since what studies often find to precede actions appears as an intuitive, automatic process or impulse, it has been claimed that reason is not a cause of actions at all, that humans are not reasonable beings, or that no such thing as reason, understood as a process of deliberation, actually

and/or behaviour in such cases are "implicit" means that they are typically not ascribable agentially: People are usually not aware of them and will even provide "post-hoc" reasons to justify their biased conclusions (compare fn. 2). Since our behaviour is thereby revealed to be at least partially explained by causes which are not reasons, it seems that we might have to deny, firstly, that actions are caused by reasons, and, secondly, that psychological laws rely on reasons as causes.

Now, the second point should be partially conceded, leading to a clarification of the notion of "psychological explanation": When they are not concerned with the explanation of behaviour which is under agential control and explainable by the ascription of intentional attitudes, psychological laws need indeed not invoke rational relations: They simply need to capture how thoughts and/or behaviour depend on internal and/or external conditions, and since agents may well behave systematically (i.e. in a specifiable and generalizable way) irrationally in the sense implied by research about implicit biases, there can turn out to be psychological laws of irrationality (cf. Ariely 2010). Of course it is simply wrong to require psychological explanation tout court to have to rely on reason explanations; that much should already be clear from even superficially browsing the current psychological literature. However, it is worth noting that the stated psychological effects are only describable as irrational when contrasted with an appropriate rational norm. Therefore, not only does rationalization constitute a methodological requirement for mental state attribution, it also supplies a foil for singling out biases as such, namely as deviations from what is normatively required or desired (cf. Davidson 2004, 180).

The first point, however – that actions are caused by reasons –, is not falsified merely by the existence of biases and similar confounders of rational cognition and behaviour. Their mere existence does not show that reasons are not generally causally effective and that reasons do not make for valuable predictors of behaviour. Even establishing that biases have the potential to override rational reasoning processes is not enough for inferring that reasons are generally ineffective and that reason-explanations are invaluable (cf. Triskiel 2016, 88 f.). Indeed, it would be irresponsible to

exists. Obviously, nothing in this paper called reason needs to be operationalized in this way (cf. Sauer 2012; 2017).

disregard reasons as commonplace determinants of behaviour: For example, knowing that Smith believes that boarding for his flight is about to start at gate 7 enables you to predict, ceteris paribus, where Smith is going to be next. Insights about biases, as valuable as they are for adjusting and correcting biased cognition and behaviour, cannot possibly undermine this kind of explanation in general (compare Fodor 1989, chap. 1). Rather, what research about biases shows is that they make for *additional* psychological causes beside reasons. And whenever intentional explanation remains valuable, methodological rationalization needs to be applicable. Hence, insights about biases can be seen as restricting the present domain of inquiry: They reveal the conditions under which identifying psychological causes does not require rationalization in Davidson's sense.

Now, it is still not entirely clear what the kind of rationalization required for intentional explanation exactly amounts to. While the claim that rationality must be ascribed methodologically indeed suggests a conflict with the possibility of irrationality running rampant, it is unclear how severe that conflict must be. For instance, just how irrational can people be without becoming intentionally unexplainable? Davidson clearly thought that there is a trade-off between the ascription of irrationality and intentional explainability (see Davidson 2001a, 159, as quoted at length above), but he would not draw a line (indeed, given his and Quine's stance on the indeterminacy of translation and interpretation, we should assume that, according to this position, there is no specific line, but that there are indefinitely many potential lines). But there are many further pertinent questions: If for every agent at any given time there is a maximally rational intentional description of her thoughts and actions available, under what conditions should we stray from it in our actual description of her thoughts and actions? How far could or should we stray from it? Should all of her mental states and actions exert the same "rationalization pressure" on this description, or are there "tentpoles" or minimal requirements (cf. Cherniak 1981) – i.e. restricted sets of mental states and/or actions whose consistency takes precedence over those excluded from these sets? If so, which are these, and to what degree(s) do they take precedence? And so on.

Given the possibility of answering these questions differently, and the added possibility of combining different answers, the possible routes to what can be viewed as an adequate kind of rationalization methodology multiply and diverge rapidly. Since I cannot consider all of these views, I

will fashion my following argument so as to address the strictest possible rationalization methodology. That is, what I will point out in the following will serve as an argument for establishing that even the strongest possible form of rationalization does not conflict with the possibility of irrationality running rampant. If this argument succeeds, then it can be ruled out that *any* methodological requirements of intentional explanation conflict with irrationality running rampant. This strongest possible form of rationalization is the following: If a person A thinks P or does Q, where P is an intentional mental state and Q is an intentional action, then it follows (just from the methodology employed in order to determine that A thinks or does P) that A's mental states are consistent, that holding P is rational in relation to A's other mental states, and that the best explanation of Q is a logical implication of P's pertinent instrumental belief(s) and desire(s). Call this form "radical rationalization". It is radical because it makes irrationality, as far as the concept pertains to the relations between A's mental states, as well as between her mental states and actions, conceptually impossible: Whatever A thinks or does is always the most reasonable thing to think or do from her point of view.

## 3. Two kinds of rationality

While it does not coincide with the distinction between the causal and rational aspects of intentional explanation, the dichotomy we are about to explore is rooted in it. As pointed out in the previous section, intentional explanations have to pick out a reason that is or was efficacious in the agent's mind – one that could cause her action by being part of her mindset (or "mindware"; Perkins 1995; Stanovich 2009). This notion of efficacy does not require an immediate awareness or subjective transparency of one's reasons, but marks a kind of reason which is an efficacious psychological cause. Any kind of unconscious or implicit cause can qualify as well. This notion is merely to be distinguished from reasons which are not present in an accurate description of an agent's mind, and/or which are psychologically inefficacious (relative to the individual and the moment which the description refers to).

I will define psychological efficacy as follows: A reason R is psychologically efficacious iff (1) R belongs to a person P's mind, (2) R has (direct

or indirect) causal powers in regard to P's actions, and (3) R's potential, anticipated or actual causal effects are, under a relevant description, logically consistent with R. Some clarifying remarks: (1) effectively distinguishes psychologically efficacious reasons from mind-external reasons by which an agent could also be judged irrational, thereby providing two distinct grounds for attributions of irrationality; (2) establishes causality between reasons and actions, and (3) establishes the rationality relation between reasons and actions. The descriptions mentioned in (3) are relevant if they manage to plausibly relate the respective effect to the agent's intentions (if only by construing the effect as a deviation from what was intended). (2) and (3) jointly allow psychologically efficacious reasons to serve their characteristic role in intentional explanation (that is, (2) and (3) explicate what it means to figure in a practical syllogism).

Psychological efficacy, so construed, pertains only to "internal components", if you will, of a given subject's mind. As mentioned, this construal takes a hint from Stanovich's (2009, 129) notion of *mindware*, which he defines as the totality of the memory-stored entities guiding decision-making and problem-solving. Like my notion of psychological efficacy, his construal marks things which (by way of being mentally stored and retrievable) belong to individual agents' minds and can causally affect their actions and/or their other mental states. Accordingly, psychologically efficacious reasons contrast with reasons which lack such efficacy because they either do not belong to P's mind (such as reasons P is oblivious to; cf. Stanovich 2009), or because they are dismissed by P (e.g. due to their being judged invalid and/or irrelevant).[8] Since, trivially, intentional interpretation is only concerned with ascribing those states to a person which belong to her own mind, it is reconcilable with the kind of irrationality that stems from acting against reasons which are rational but either external to her mind or which she dismisses.

There are, of course, other reasons apart from those stored in an agent's mind which can causally affect her actions and/or her other mental states. For instance, those given by other minds can do so. If I am told to do something for a reason I had not been aware of, but whose justification I agree

[8] Accordingly, reasons become psychologically efficacious by (1) the agent being made aware of them and/or (2) the agent making them part of the respective reasoning process (e.g. by judging them to be relevant for the reasoning process at hand).

with, it can causally affect my mind and/or behaviour, without the content of what I am being told needing to be recalled from my own memory. Accordingly, my use of the word "internal" simply means to mark reasons which mentally belong to the agent in question. This "internalism" is not to be understood as contradicting philosophical externalisms, such as semantic externalism (cf. Burge 1979) or the "extended mind" view (cf. Clark & Chalmers 1998). That is, mental content may well be individuated externally but still be had by individual agents; and bodily external but cognitively seamlessly accessible storage devices (such as notebooks; cf. ibid.) can qualify as belonging to individual minds – thereby "extending" them – and hence as in this sense "internal" as well. What is mentally external in the presently relevant sense is what belongs to other minds, to other persons. Hence, the distinction between psychologically efficacious and inefficacious reasons is based on that between the mental content of a subject's mind and the mental content external to this subject's mind. The former is (potentially) directly psychologically efficacious, the latter can only be indirectly efficacious by way of communication, transmission, or access to other minds. While all reasons which are acted upon are psychologically efficacious, not every reason by which we assess a given person to be rational or irrational is a psychologically efficacious reason. And since only psychologically efficacious reasons are relevant for intentional interpretation, there is room for a kind of irrationality which does not touch interpretability.

Two immediate ways in which reasons can be inefficacious is when an agent is oblivious to them or when she cannot grasp them. In such cases, reasons cannot even subconsciously or implicitly provide the kind of generalizable explanations which is characteristic for them (which is not to say they cannot have *any* effect on an agent who fails to grasp them; being aware of not grasping them can still have a frustrating effect, or the like). Since by far not all reasons are available to everyone, their being rational alone cannot make them agentially explanatory. So, an agent's acting on reasons can only ever be rational if they are available to her. If they are not, then her not acting upon them cannot be irrational, and neither need any of her other mental states, which would contradict her potentially doing so, be irrational, since they can only be irrational in relation to cognitively available reasons. (Notably, her ignorance itself constitutes an intentional state which explains why the respective reason

fails to be efficacious – just as anyone's holding two inconsistent beliefs can be explained by their being unaware of the way they contradict each other. For example, most of us can readily both believe that Elton John sang "Crocodile Rock" and doubt that Reginald Kenneth Dwight sang "Crocodile Rock", namely when being ignorant of the fact that both names designate the same person.)

When it comes to the methodological requirements of intentional explanation, an agent's being rational or irrational are not a matter of acting upon or considering everything that is generally rational, everything specified by good reasons (cf. McNaughton & Rawling 2004, 126). Hence, the notion of "all things considered" (cf. Davidson 1980, 21-42), which is notorious in debates revolving around irrationality, does, of course, not mean that *all* things are considered, only that all cognitively available things are. The kind of rationality which matters for intentional explanation is that which is attributed relative to an agent's mindset and actions. Explanation by intentional states is concerned with psychological efficacy; and what is psychologically efficacious does not coincide with what is rational tout court. As long as what we mean by "reason" is an explanatorily valuable psychological cause, it is implied that this reason is an available, efficacious part of an agent's mindset.

When I introduced the notion of radical rationalization in the previous section, one might have objected that, given our cognitive limitations, the empirical knowledge we have about human psychology, as well as the practical constraints of everyday requirements, no such picture should assume that intentional explanation requires a maximally rational interpretation of a maximal set of pertinent evidence at any given moment. Even if it were possible for cognitively limited agents like us to perform such interpretations, it would be highly inefficient. Rather, we undoubtedly assume that non-pathological agents use and understand standard senses of the terms belonging to the language they are speaking, that their understanding is somewhat consistent over time, that they are generally interpretable using the same (or similar enough) function mapping actions and contexts to intentional states, and so on.

Consequently, even if radical rationalization should rule out irrationality, failures of rationality may creep back in, namely as markers of diachronic deviations in meaning, deviations from norms of word use, deviations from what actions are standardly performed in a given context, and

so on. But, given what I pointed out just now, these forms of irrationality can also be sharply distinguished from failures of psychological efficacy. This is because, when in doubt about an agent's psychological causes, methodological rationalization requires us to go ahead and assign deviant understandings or deviant mental states as psychological causes. This deviance is a deviance from an external norm (and often merely from a pragmatic one) and has no bearing on that person's intentional explainability. For example, we may very well go ahead and stipulate that a given person means "yawl" by the expression "ketch" in order to arrive at a sufficiently rational interpretation of her mental states and behaviour (cf. Davidson 2001a, 196). Here, she deviates from the external norm that yawls should be referred to by the word "yawl". If her behaviour regarding yawls is irrational, it is only to the extent that she violates this external norm, but certainly not in the sense that it exhibits a failure of psychological efficacy. Psychological efficacy is satisfied insofar as her describing a yawl as a "ketch" is psychologically motivated by the perception of what she may very well believe to be a yawl, but *also* mistakenly believes to be referred to by the word "ketch". It is this interpretational ascription of mistakenly violating an external norm, and hence the ascription of a deviant mental state – one which deviates from an external norm – which renders her sufficiently rational.

This brings us to our second sense of the term "reason", namely that which refers to what is expressed by a psychologically external norm. Unlike our first kind, this kind of reason is external to an agent's mindset. This is not to say that such reasons cannot be psychologically efficacious, only that they are not in the case of the agent who is in this second sense irrational. As examples, consider cases in which an agent does something she might be persuaded not to do if she were made aware of its negative consequences. Strikingly, this second kind of reason can as well be internal *and* psychologically inefficacious, as in the case of a smoker's acknowledging that reason demands that ("most reasonably" or "all things considered") one should stop smoking, without at the same time taking this demand as a reason (i.e. psychological cause) for herself to stop smoking. To clearly bring out the difference between these two notions of "reason", note that there is generally no contradiction in specifying an unreasonable (i.e. normatively or logically unsound) desire as being the reason (i.e. the psychological cause) for someone's action. And the only

way we can take this real possibility to not express a contradiction is by acknowledging that these two uses of the word "reason" express two different meanings.

Even though their attribution may also have normative aspects (i.e. agents should follow *some* norms of rationality in order to be able to have reasons at all), psychological causes must, in any robustly empirical psychological theory, be descriptive notions: They must not be attributable on the grounds of logical ("a priori") criteria alone, but on descriptive ones also, such as by whether they are part of the actual psychological make-up of an agent.[9]

Again, this is not to say that what is rational in a second sense – in the sense that goes beyond what descriptively persists in an agent's mind – cannot be psychologically efficacious, but that it can be so only by becoming part of an agent's mindset. That is, reasons beyond our own mindset (from now on referred to as "reasons$_2$" which are "rational$_2$") potentially shape our minds and our actions insofar as we have the means and the reasons in our first sense (from now on referred to as "reasons$_1$" which are "rational$_1$") to act in accordance with these reasons$_2$. That is, in order for mind-external reasons$_2$ to become mind-internal reasons$_1$, they need to be both accessible to our minds and there need to be some reasons$_1$ motivating their incorporation. For example, perhaps there is someone who desires apples more than oranges, and oranges more than bananas, but bananas more than apples. Any such preference ordering $A > B > C > A$ is irrational in the sense that it makes us exploitable: on a behavioural interpretation, it means that we are willing to trade A and some sum for B, then to trade B and some sum for C, then C and some sum for A, ad infinitum, thus losing everything while never gaining anything – anything but the satisfaction of our irrational desire, perhaps (cf. Ramsey 1931, 156-198; Davidson et al. 1955). Yet, while they are in this sense irrational, it may be true of anyone

[9] Although I am willing to consider that intentional explanation has both normative and descriptive aspects, I will not take a stance on whether it could not rather be *purely* descriptive or empirical (insofar as there might be purely descriptive explication of what it means to follow the kinds of norms of rationality necessary for having reasons). I only mean to say that it must at least be *also* empirical in order to be (quasi-)scientific. This robustness criterion is inspired by Piccinini (2007).

that they have these desires, and in such a case these causally explain their actions.

Firstly, such cases illustrate that irrational$_2$ mental states can constitute psychological causes (because these only need to be rational$_1$). Accordingly, Davidson did not take believing in astrology, flying saucers or witches, intending to climb Mount Everest without oxygen, or trying to square the circle as necessarily irrational (Davidson 2004, 170).[10] (Of course, we should add that they *can* also be irrational$_1$, depending on whether they contradict the agent's *other* relevant mental states.) Yet, anyone who has these irrational desires may wise up as a consequence of being made aware of their exploitability and cease acting on such desires. Thereby, such cases secondly illustrate that norms of rationality$_2$ can become reasons$_1$ by becoming part of an agent's mindset. Such norms do not *directly* enter into predictions or explanations of intentional psychology, but only by way of their psychological efficacy. Hence, even if we both assume that intentional explanation requires radical rationalization and that irrationality$_2$ may run rampant, people can still be explainable intentionally.

I wish to point out one last kind of cases when the two kinds of irrationality can come apart, one that is perhaps more interesting than ignorance and failure to grasp the rational validity or content of a norm. In the latter cases, norms cannot be psychologically efficacious because they are external to an agent's mindset. But, akrasia aside (see fn. 4), how can it be the case that reasons are cognitively available to an agent, yet still fail to be efficacious? One way we can construe this possibility is to consider a kind

---

[10] However, we can also find Davidson expressing that he does "not think we can clearly say what should convince us that a man at a given time (or without any change of mind) preferred a to b, b to c, and c to a" (Davidson 1980, 237). This might be read in the following way: Perhaps the principle of transitivity is so fundamental that attributing its application to an agent's mental states constitutes a conditio sine qua non. This, in turn, would suggest that Davidson's holism is restricted: If there are mental states sine qua non, then not every mental state attribution depends on other mental states. I will not pursue this point further, but only suggest that holism might be defended by considering cases in which, other things being equal, attributing this principle's violation to an agent makes her appear more rational and is thereby justified. (For the question whether there are principles which constitute conditions sine qua non, see Cherniak 1981.)

of partiality on behalf of the agent. That is, agents can fail to abide by norms whose abiding by is, ceteris paribus, generally found to be rational if they are partial to not abiding by them. Since there can be no general impartiality (and since it need not be generally rational to be partial to a certain cause), a given norm may be judged perfectly rational, while abiding by it is not necessary for a specific agent to be rational. In such cases, we should assume that the agent in question has a set of psychologically efficacious reasons which motivate her to dismiss the external norm by which she might be judged irrational.

Returning to our previous example, someone's action of smoking could be judged irrational if she has sufficient reason to establish the general principle that one shouldn't smoke. (One need not be aware of medical details regarding the consequences of smoking here; vaguely knowing that smoking is unhealthy may already suffice to establish this principle.) In other words, if one has good reason for believing this principle to be rational, but still smokes, then one is irrational$_2$. Why does this constitute irrationality$_2$ rather than irrationality$_1$? Why are the respective reasons not psychologically efficacious, even though they are part of the agent's mindset? It is because the psychological inefficacy of a generally rational principle can be accommodated by intentional explanations by way of attributing partiality to an opposing principle to the agent (such as an overriding desire to smoke), thereby establishing rational$_1$ consistency. Of course, any such partiality (e.g. any desire of this kind) must not be assumed in the face of sufficient evidence against it; but any such evidence must in turn be weighed against the evidence of the agent's smoking in the face of realizing that, generally, one shouldn't smoke.[11]

[11] This point may be likened to Davidson's, that reasons of the "all things considered"-type (abbreviated as ATC) may still fail to make for reasons to act, or all-out reasons (abbreviated as AO; cf. Davidson 1980, pp. 21-42). However, I take my point to be more clearly committed to construing the eventual AO as the superior reason to the ATC. According to radical rationalization, we are committed to thinking of the ATC as an external norm: One which the agent thinks should "rationally$_2$" be imposed on her, but which she rejects because of her partiality to the AO. Taken literally, there can be no such thing as overriding an internal ATC, since, if all (available) reasons have truly been considered, then the strongest of them must be an AO. According to radical rationalization, it is impossible to decide against one's best reason. One may, of course, still deceive oneself by thinking that what one did was not for

Finally, I would like to address a potential objection, namely that the distinction between rationality$_1$ and rationality$_2$ is so vague as not to be helpful. This objection is rooted in the already briefly mentioned worry that, given Davidsonian holism (cf. Davidson 1980, 256 f.; Davidson 2001a, 22, 200; Davidson 2001b, 98), there is no specific list of what mental states a person must have in order to count as rational$_1$. One counter to this objection consists in the delineation of a substantial "minimal rationality" (cf. Cherniak 1981). According to this view, in order to be interpretable as agents, persons need only satisfy such minimal criteria of rationality, while the mental states or actions not satisfying these criteria could play the role of being rational$_2$. Providing criteria for what is minimally rational ultimately makes for a substantial list of mental states sufficient for agential interpretation. Adopting this view may in turn require a modification of Davidson's view, since a plausible interpretation of Davidson's holism – i.e. that there are no specific mental states which a person must hold in order to be interpretable as an agent, but only a sufficient amount – could even be taken to speak against the substantial minimal rationality view by finding that some (or even *any*) mental state(s), holding which is necessary for a given person to be interpretable, could figure as being unnecessary for interpreting a different person. In other words, Davidsonian holism allows that, for any set of mental states which the substantial minimal rationality view might construe as being necessarily had by any agent in order for her to be interpretable, there is a person who is interpretable but does not have these mental states.

In any case, substantial delineations of minimal rationality are in fact unnecessary to defend our present distinction between rationality$_1$ and rationality$_2$ against the objection that it would be undermined by not being able to neatly sort mental states or actions into one or another category. That is, the fact that we might not be able to generally sort each mental state into one of the two categories opened up by our distinction does not mean there cannot be a distinction. In order to make the distinction, it is already sufficient that there is a set of mental states for each given person which is necessary for interpreting her as an agent (this is the set which is

---

the best reason one had – but it really only shows that one does not think of the best external reason as making for the best internal one (in a sense which might not be subjectively transparent).

"internal", "cognitively accessible", part of her "mindware", etc. – it is the set which is psychologically efficacious), whereas the rest is unnecessary (whose content corresponds to a second set of "external" reasons, which may be rational$_2$). So, even if a substantial "minimal rationality" could not be defended as consistent with Davidsonian holism, our distinction would remain solid.

## 4. Conclusion

I have argued that the form of rationality methodologically required for intentional explanation comes apart from a common understanding of rationality. What the former requires is for an agent to be consistent both in terms of her mental states and in terms of her behaviour. On the other hand, the idea that human beings are notoriously irrational creatures – an idea which is both popular and also seems to be supported by recent psychological research – is grounded in the perception that people act against plausible external norms, not that there couldn't possibly be specifiable reasons for their actions. Specifically, I have argued that even the strongest form of methodological rationalization (which I called "radical rationalization") does not prohibit irrationality in this second sense – simply because the criteria for irrationality in the second sense, namely violating external norms of rationality, are independent of how strongly we phrase our criteria for rationality in the first sense. Whether one needs to be absolutely internally consistent or just more or less internally consistent makes no difference for whether the mental states attributed on the basis of achieving this consistency (can) violate some external norm of rationality. In other words, even on the strongest assumption – that perfect internal consistency already follows from being interpretable –, inconsistency with some external norm is nonetheless possible. Calling someone irrational in this second sense is akin to calling out their reasons as short-sighted, foolish or immoral, not as inconsistent in terms of the methodology required for attributing intentional explanation. While the first kind of rationality, by being connected to methodological rationalization, persists in internal consistency, the second kind does not.

Consequently, we can distinguish between two different uses of the term "rational": We may say that if Anne were rational, she would vote for

the Green Party, and mean that she should do so in order for her actions to reach a maximum of consistency with her mental states (such as her beliefs about environmental policies). This falls squarely into the domain of intentional ascriptions, for we would strive to find a way to make her voting behaviour consistent with her beliefs. Or we can mean that if her actions were most consistent with what we perceive to be the state of the world, she should vote Green. We might perceive the state of the world to be one of threatening global warming and waning natural resources, and consequently think it only right to vote Green – even if none of Anne's beliefs are actually consistent with voting Green (as she might steadfastly deny global warming and the waning of natural resources).

Reconsidering our opening quotes, it should be clear that the lamented forms of irrationality are not those methodologically excluded by intentional explanation. Neither Russell nor Stanovich meant to imply that human beings generally fail to weep over a loved one's death, fail to value acts of kindness, fail to eat when hungry and presented with food, or fail to understand why insurgents would not freely surrender their children to the oppressor. That is, they do not mean to criticize people for failing to understand the basic concepts with which we describe their mental states and/or actions. Rather, both mean to criticize failures of critical and long-term thinking and the like: a lack of adherence to reasonable norms. Criticism of this kind does not touch the fact that we have to assume an agent to be rational in order to render her thoughts and actions intelligible, and that, barring pathological instances, we can expect to be able to generally interpret them in this way.

## Acknowledgements

I would like to thank Steffen Steinert, Ali Yousefi, Mario Günther, Stephan Sellmaier, and two anonymous referees for their helpful comments on an earlier version of this paper.

## References

ARIELY, D. (2010): *Predictably Irrational, Revised and Expanded Edition: The Hidden Forces That Shape Our Decisions*. New York: Harper Perennial.

ARPALY, N. (2000): On Acting Rationally against One's Better Judgment. *Ethics* 110, 488-513.

BECHTEL, W., & WRIGHT, C. (2009): What is Psychological Explanation? In: Calvo, P. & Symons, J. (eds.): *Routledge Companion to Philosophy of Psychology*. London: Routledge, 113-130.

BRATMAN, M. (1979): Practical Reasoning and Weakness of the Will. *Noûs* 13, 153-171.

BROADIE, A. (1968): The Practical Syllogism. *Analysis* 29, No. 1, 26-28.

BURGE, T. (1979): Individualism and the Mental. *Midwest Studies in Philosophy* 4, No. 1, 73-122.

BUSS, S. (1997): Weakness of Will. *Pacific Philosophical Quarterly* 78, 13-44.

CHERNIAK, C. (1981): Minimal Rationality. *Mind* 90, No. 358, 161-183.

CLARK, A., & CHALMERS, D. (1998): The Extended Mind. *Analysis* 58, No. 1, 7-19.

CUMMINS, R. (2000): "How Does It Work" vs "What Are the Laws?": Two Conceptions of Psychological Explanation. In: Keil, F. & Wilson, R. (eds.): *Explanation and Cognition*. Cambridge (Mass.): MIT Press, 117-144.

DAVIDSON, D. (1980): *Essays on Actions and Events*. Oxford: Clarendon Press.

DAVIDSON, D. (2001a): *Inquiries into Truth and Interpretation*. 2nd ed. Oxford: Clarendon Press.

DAVIDSON, D. (2001b): *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press.

DAVIDSON, D. (2004): *Problems of Rationality*. Oxford: Clarendon Press.

DAVIDSON, D., MCKINSEY, J. & SUPPES, P. (1955): Outlines of a Formal Theory of Value. *Philosophy of Science* 22, 140-160.

DENNETT, D. (1987): *The Intentional Stance*. Cambridge (Mass.): MIT Press.

DENNETT, D. (1991): Real Patterns. *The Journal of Philosophy* 88, No. 1, 27-51.

FODOR, J. (1989): *Psychosemantics*. Cambridge (Mass.): MIT Press.

GIGERENZER, G. (2008): Moral Intuition = Fast and Frugal Heuristics? In: Sinnott-Armstrong, W. (ed.): *Moral Psychology, Vol. 2: The Cognitive Science of Morality: Intuition and Diversity*. Cambridge (Mass.): MIT Press, 1-26.

HEMPEL, C., & OPPENHEIM, P. (1948): Studies in the Logic of Explanation. *Philosophy of Science* 15, 135-175.

HOLTON, R. (1999): Intention and Weakness of Will. *Journal of Philosophy* 96, 241-262.

LEWIS, D. (1983): *Philosophical Papers*. Vol. 1. Oxford: Oxford University Press.

MCNAUGHTON, D. & RAWLING, P. (2004): Duty, Rationality, and Practical Reasons. In: Rawling; P. & Mele, A. (eds.): *The Oxford Handbook of Rationality*. Oxford: Oxford University Press, 110-131.

MELE, A. (1987): *Irrationality*. New York: Oxford University Press.

PERKINS, D. (1995): *Outsmarting IQ: The Emerging Science of Learnable Intelligence*. New York: Free Press.

PICCININI, G. (2007): Computational Explanation and Mechanistic Explanation of Mind. In: de Caro, M., Ferretti, F. & Marraffa, M. (eds.): *Cartographies of the Mind: The Interface between Philosophy and Cognitive Science*. Dordrecht: Springer, 23-36.

RAMSEY, F. (1931): *The Foundations of Mathematics and Other Logical Essays*. Braithwaite, R. (ed.), London: Routledge & Kegan Paul, Trench, Trubner & Co. – New York: Harcourt, Brace and Company.

RUSSELL, B. (1950): *Unpopular Essays*. London – New York: Routledge.

SAUER, H. (2012): Educated Intuitions. Automaticity and Rationality in Moral Judgment. *Philosophical Explorations* 15, No. 3, 255-275.

SAUER, H. (2017): *Moral Judgments as Educated Intuitions*. Cambridge (Mass.): MIT Press.

SEARLE, J. (1979): What is an Intentional State? *Mind* 88, No. 1, 74-92.

SEARLE, J. (2000): *Mind, Language and Society: Philosophy in the Real World*. London: Phoenix.

SELLARS, W. (1997): *Empiricism and the Philosophy of Mind*. Cambridge (Mass.) – London: Harvard University Press.

SIE, M. & WOUTERS, A. (2010): The BCN Challenge to Compatibilist Free Will and Personal Responsibility. *Neuroethics* 3, 121-133.

STANOVICH, K. (2003): The Fundamental Computational Biases of Human Cognition: Heuristics That (Sometimes) Impair Decision Making and Problem Solving. In: Davidson, J. & Sternberg, R. (eds.): *The Psychology of Problem Solving*. New York: Cambridge University Press, 291-342.

STANOVICH, K. (2009): *What Intelligence Tests Miss: The Psychology of Rational Thought*. New Haven: Yale University Press.

SUNSTEIN, C. (2005): Moral Heuristics. *Behavioral and Brain Sciences* 28, No. 4, 531-573.

TRISKIEL, J. (2016): Psychology Instead of Ethics? Why Psychological Research Is Important but Cannot Replace Ethics. In: Brand, C. (ed.): *Dual-Process Theories in Moral Psychology*. Wiesbaden: Springer, 77-98.

TVERSKY, A. & KAHNEMAN, D. (1974): Judgment under Uncertainty: Heuristics and Biases. *Science* 185, No. 4157, 1124-1131.