Incoherence and Irrationality

by Donald Davidson•

Summary

To judge a belief, emotion, or action irrational is to make a normative judgment. Can such judgments be objective? It is argued that in an important class of cases they can be. The cases are those in which a person has a set of attitudes which are inconsistent by his or her own standards, and those standards are constitutive of the attitudes. Constitutive standards are standards with which an agents' attitudes and intentional actions must generally accord if judgments of irrationality are to be intelligible.

Résumé

Juger irrationnelles une croyance, une émotion ou une action, c'est faire un jugement normatif. De tels jugements peuvent-ils être objectifs? On montre qu'ils peuvent l'être dans une classe importante de cas: ceux où une personne a un ensemble d'attitudes qui sont inconsistantes avec ses propres normes, ces normes étant constitutives des attitudes considérées. Les normes constitutives sont des normes avec lesquelles les attitudes et les actions intentionnelles des agents doivent s'accorder pour que les jugements d'irrationalité deviennent intelligibles.

Zusammenfassung

Wenn wir einen Glauben, eine Emotion oder eine Handlung als irrational beurteilen, so fällen wir ein normatives Urteil. Kann ein solches objektiv sein? Es wird argumentiert, dass dies für eine wichtige Klasse von Fällen zutrifft. Die Fälle sind diejenigen, in welchen eine Person eine Menge von Haltungen hat, die nach ihren eigenen Massstäben unvereinbar sind, und wo diese Massstäbe für die Haltungen konstitutiv sind. Konstitutive Massstäbe sind solche, mit denen die Haltungen und absichtlichen Handlungen des Handelnden im allgemeinen übereinstimmen müssen, wenn Urteile über Irrationalität verständlich sein sollen.

Irrationality, like rationality, is a normative concept. Someone who acts or reasons irrationally, or whose beliefs or emotions are irrational, has departed from a standard; but what standard, or whose, is to be the judge? If you deviate from my norms of rationality, and you do not share my sense of what is reasonable, then are you really irrational? After all, fully rational agents can differ over values. If rationality is just one more value or complex set of

• The University of California, Berkeley, and Balliol College, Oxford.

values, then calling someone irrational would seem to be no more than a matter of expressing disagreement with his values or norms.

No doubt we very often stigmatize an action, belief, attitude, or piece of reasoning as irrational simply because we disapprove, disagree, are offended, or find something not up to our own standards. I am not concerned with such cases in this paper. My interest here is entirely with cases, if such there be, in which the judgment that the works or thoughts of an agent are irrational is not based, or at least not necessarily based, on disagreement over fact or norm — objective irrationality, one might be tempted to call it. This suggests that we should limit ourselves to cases in which an agent acts, thinks, or feels counter to his own conception of what is reasonable; cases where there is some sort of inner inconsistency or incoherence.

Inner inconsistency is, however, hard to describe in any detail, and harder still to explain. The difficulty in describing inner inconsistency is created by the character of the so-called propositional attitudes: belief, desire, intention, and many of the emotions. Put briefly, the problem is this: one way in which propositions are identified and distinguished one from another is by their logical properties, their place in a logical network. But then it would not seem possible to have a propositional attitude that is not rationally related to other propositional attitudes. For the propositional attitude itself, like the proposition to which it is directed, is in part identified by its logical relations to other propositional attitudes. Suppose someone discovers that his rake is missing and comes to believe on slender evidence that his neighbor has stolen it. Is he (objectively) irrational? Certainly not if he deems his evidence sufficient, and has no evidence against his suspicion. But suppose he has far better evidence against his belief than for it. Still he is not irrational unless he appreciates that the evidence he has *is* evidence against his belief, and holds that the evidence against outweighs the evidence for his belief. But does even this suffice to show he is irrational if he does not accept what Carnap called "the principle of total evidence" which counsels an agent to accept the hypothesis supported by the totality of evidence he or she has?

Here we have reached an aspect of rationality so fundamental that we cannot make sense of an agent who does not generally reason in accord with it. And so we have reached a point where the distinction between the standards of rationality of the agent himself and of his critic merge. It is an "objective", though normative, judgment that someone whose reasoning is on some occasion not in accord with the principle of total evidence has reasoned irrationally. (This claim will in the end be modified.)

The difficulty in explaining irrationality is in finding a mechanism that can be accepted as appropriate to mental processes and yet does not rationalize

what is to be explained. What makes trouble is that our normal way of explaining the formation of propositional attitudes, including intentions and intentional acts, is to state the reasons that caused the attitude or act. Thus many of Freud's explanations of apparently irrational thoughts and acts are intended to show that from the agent's point of view (enlarged to embrace unconscious elements) there were good reasons for his thinking or acting. The paradoxical consequence is that explaining irrationality necessarily employs a form of explanation which rationalizes what it explains; without the element of rationality, we refuse to accept the account as appropriate to mental phenomena. We look, or tend to look, not merely for causes and forces, but for causes that are reasons. To explain irrationality we must find a way to keep what is essential to the character of the mental — which requires preserving a background of rationality — while allowing forms of causality that depart from the norms of rationality. What is needed to explain irrationality is a mental cause of an attitude, but where the cause is not a reason for the attitude it explains.

Let me take another example; one drawn from real life, or at least from my life. One late Spring afternoon I was returning home from my work at Princeton University. It was a warm day, doors stood open. I lived in one of a row of attached houses in which faculty members were housed. I walked in the door. I was not surprised to find my neighbor's wife in the house: she and my wife often visited. But I was slightly startled when, as I settled into a chair, she offered me a drink. Nevertheless, I accepted with gratitude. While she was in the kitchen making the drink I noticed that the furniture had been rearranged, something my wife did from time to time. And then I realized the furniture had not only been rearranged, but much of it was new — or new to *me*. Real insight began when it slowly came to me that the room I was in, though identical in size and shape to the room I was familiar with, was a mirror-image of that room; stairs and fireplace had switched sides, as had the door to the kitchen. I had walked into the house next to mine.

Here is a case of gross factual error. Instead of using the evidence at hand in a natural way to support the obvious hypothesis, I somehow managed to accommodate the growing evidence against the assumption that I was in my own house by fabricating more and more absurd or far-fetched explanations. Was I being irrational in believing I was in my own house? Well, that belief by itself, however strange or odd, was surely not irrational or even foolish. But given the accumulating evidence against my belief? Of course it would have been irrational to believe I was in my own house on the basis of contrary evidence. But did I have contrary evidence? Not from *my* point of view, for I thought that my neighbor's wife was being exceptionally kind in offering me a

drink in my own house; I thought my own wife had rearranged the furniture and even introduced some new furniture. I did not so much as entertain the hypothesis that I was not in my own house, and so did not make the possibly absurd mistake of supposing my evidence supported the hypothesis I was in my own house rather than in another.

Is there a point of view from which we can make out that my belief that I was in my own house was irrational? No doubt there is. I believe, like everyone else, that when I have to invent strange explanations of what I think I see or believe I should consider alternative hypotheses. If I had adhered to my own standards of hypothesis formation, of "inference to the best explanation" as Harman calls it, I would have wondered much sooner than I did whether my assumption that I was in my own house was correct. I clung to a premature assumption far too long, and in rearranging so many beliefs (subjective probabilities), I failed to apply Quine's principle of conservation: other things being equal, change as few expectations as possible when accommodating recalcitrant appearances. So there is a clear sense in which I held a pattern of beliefs not in accord with my own best standards of rationality. I was not aware of this. Nevertheless, I was in a state of *inner inconsistency*.

Suppose that, contrary to the facts, I had asked myself whether I was in my house or in my neighbor's house, and had acknowledged that the evidence, though not absolutely conclusive, favored the hypothesis that I was in my neighbor's house. Then I would again have been in a state of inner inconsistency *provided* I held to the general principle that one ought to adjust one's degree of belief in an hypothesis to what one deems to be the extent to which it is supported all one's available evidence — what one takes to be the available evidence, of course, since one can do no better.

What this example, with its various applications, suggests is that no factual belief *by itself*, no matter how egregious it seems to others, can be held to be irrational. It is only when beliefs are inconsistent with other beliefs according to principles held by the agent himself — in other words, only when there is an inner inconsistency — that there is a clear case of irrationality. Strictly speaking, then, the irrationality consists not in any particular belief but in inconsistency within a set of beliefs (or within a set consisting of beliefs combined with principles, if principles are to be distinguished from beliefs). I think we must say much the same about intentions, intentional actions, and other propositional attitudes (usually, or perhaps always, in conjunction with beliefs or principles). They are never irrational in themselves, but only as part of a larger pattern.

We often do say of a single belief or action or emotion that it is irrational, but I think that on reflection it will be found that this is because we assume in

these cases that there must be an inner inconsistency. The item we choose to call irrational is apt then to be the one by rejecting which things are most easily or economically brought back into line. If I buy a lottery ticket believing it will win, you may well be right in calling me irrational. But the belief that the ticket is a winning ticket cannot by itself make me irrational; after all, I might have, or legitimately think I have, inside information. In accusing me of irrationality you assume I have no such information; you assume I know I have only one chance in many of winning, and in the light of *this* belief (and further beliefs and principles), my belief that I will win is absurd. My beliefs cannot be made to fit together according to my own views of how probabilities should be distributed over beliefs.

Or suppose that I am ashamed that I am not six feet tall. Such an emotion is, many would hold, an irrational emotion. If it is irrational, the reason must be something like this: one can be ashamed of having some trait only if one believes one has it and holds that having that trait is blameworthy. But something is blameworthy only if it is something for which one is responsible, and one cannot be responsible for not being six feet tall. If something like this account is correct, we again find that irrationality is a feature of a complex of attitudes, not of isolated parts of the complex. It may be that I think I *am* responsible for my not being six feet tall; then I am not, after all, irrational in being ashamed of not being six feet tall. Of course, my belief that I am responsible for my not being six feet tall may itself be inconsistent with other things I believe, in which case irrationality is present in another way. The point remains: we call a single attitude, belief, or action irrational only when we assume it conflicts with other beliefs or attitudes of the agent.

Here is an example of an irrational *action*. I stay up late arguing with a friend about politics even though I know I will not be able to change his mind (nor he mine) and I do not enjoy the clash of opinion. My action is an example of akrasia, since I am acting contrary to my own best judgment. No doubt there are reasons why I go on arguing: I am exasperated by my friend's false views and warped values (as I see them), and I cannot resist the desire to set him straight, even though I know I will not succeed. I have my reasons for acting as I do, then, but these reasons are outweighed, in my own sober judgment, by the reasons I have against continuing the argument. Once more, it is not the isolated item, in this case the action itself, that proves irrationality. The irrationality depends on the discrepancy between the action and the reasons I recognize as relevant to its performance.

As in the other examples, there is much more to say here, and the need for distinctions. Any action, for example, may be described in endless ways that are irrelevant to its irrationality even in context. But it is always relevant to

questions of rationality and irrationality to consider the description of an action under which it is intentional. An intention (or so I have argued)[1] consists in an evaluative judgment of a certain kind, and in the case of my ill-considered late night political argument, this judgement is literally inconsistent with the judgement enjoined by the "principle of continence" which says one should prefer (act on) the judgment based on all the considerations deemed relevant. Intentional actions entail the existence of intentions, and so acting with a certain intention can entail the existence of a judgment that is inconsistent with other attitudes and principles of the agent. Strictly speaking, then, we might want to say the irrationality lies in the inconsistency of the intention with other attitudes and principles rather than in the inconsistency of the action of which it is an intention with those attitudes and principles.

So far, my thesis (far from proven, of course) is that all (objective) irrationality is a matter of inner inconsistency. But there is a difficulty which brings us back to the question with which I began: what, or whose, standards are at stake? It may seem that this matter was settled when it was decided that irrationality is always inner; this might be taken to show that the standards that matter are those of the agent alone. However, here there lurks an unexplored and undefended assumption which might be put this way: why must *inconsistency* be considered irrational? (Alternatively, or perhaps equivalently, one could ask: who is to decide what consistency demands?) Isn't this just one more evaluative judgment, and one that an agent might reject? Emerson did not see consistency as an intellectual virtue. When sufficiently aroused, my father would sometimes reply to the accusation that he had contradicted himself by saying, "I'll contradict myself if I want to."

Let me take one more example. Imagine that you want to rent a house, and three houses are available, a large house that rents for $1,000 a month, a medium-sized house that rents for $800 a month, and a small house that rents for $600 a month. You prefer the large house to the medium-sized house, since the difference in cost is relatively small; you prefer the medium to the small on the same ground. But you also prefer the small to the large, since in this case the difference in cost is enough to outweigh considerations of size. Is the set of your preferences irrational? I may remind you that according to rational decision theory your preferences form an inconsistant triad, and so you are irrational. Suppose you reply, "So what; those are *your* standards of rationality, not mine." "Well (I argue), decision theory (and common sense) says to choose an option available to you such that none is preferred to it.

[1] In Essay 5 in *Essays on Actions and Events*, Oxford University Press, 1980. See also my replies to Michael Bratman, Paul Grice and Judith Baker, and Christopher Peacocke in *Essays on Davidson, Actions and Events*, Bruce Vermazen and Merrill Hintikka (eds.), Oxford University Press, 1985.

How can you do this, since whatever option you hit on, there is another you like better?" "Hold on (you retort), what *are* my options? If they are the large house and the medium, I take the large; if the medium and the small, I take the medium; if the large and the small, I take the small." "Aha! (I snort) And suppose I offer you all three; then what?" "Easy (you smile): I take the large." "But you prefer the small to the large." "Only (you reply) in case my choice is between the large and the small only; if the medium is also available, I prefer the large."

At this point there are several lines I might take. I might complain that it is irrational to change one's preference of the large over the small just because another option is available; but I may have trouble explaining why this is irrational. Or I may point out that a dutch book can be made against you: given your declared preferences, you can be offered a set of bets such that no matter what happens you lose by your own admission. Plenty of questionable assumptions are needed for this argument.

I am strongly inclined to think my mistake in this imagined exchange came right at the start: I should never have tried to pin you down to an admission that you ought to subscribe to the principles of decision theory. For I think everyone does subscribe to those principles, whether he knows it or not. This does not imply, of course, that no one ever reasons, believes, chooses, or acts contrary to those principles, but only that if someone does go against those principles, he goes against his own principles.

I would say the same about the basic principles of logic, the principle of total evidence for inductive reasoning, or the analogous principle of continence. These are principles shared by all creatures that have propositional attitudes or act intentionally; and since I am (I hope) one of those creatures, I can put it this way: all thinking creatures subscribe to *my* basic standards or norms of rationality. This sounds sweeping, even authoritarian, but it comes to no more than this, that it is a condition of having thoughts, judgments, and intentions that the basic standards of rationality have application. The reason is this. Beliefs, intentions, and desires are identified, first, by their causal relations to events and objects in the world, and, second, by their relations to one another. A belief that it is about to rain would lose much of its claim to be just *that* belief if it did not have some tendency to cause someone who had it and wanted to stay dry to take appropriate action, such as carrying an umbrella. Nor would a belief that it is about to rain plausibly be identified as such if someone who was thought to have that belief also believed that if it rains it pours and did not believe it was about to pour. And so on: these obvious logical relations amongst beliefs; amongst beliefs, desires and intentions; between beliefs and the world, make beliefs the beliefs they are; there-

fore they cannot in general lose these relations and remain the same beliefs. Such relations are *constitutive* of the propositional attitudes.

I have greatly oversimplified by making it seem that there is a definite, and short, list of "basic principles of rationality". There is no such list. The kinds and degrees of deviation from the norms of rationality that we can understand or explain are not settled in advance. We make sense of aberrations when they are seen against a background of rationality; but the background can be constituted in various ways to make various forms of battiness comprehensible. So it would be a mistake to put too much weight on the examples of irrationality that I have chosen, and worse to worry whether I have in each case drawn the line between principles constitutive of rationality and potentially intelligible flaws in just the right place. The essential point is that the more flamboyant the irrationality we ascribe to an agent, the less clear it is how to describe any of his attitudes, whether deviant or not, and that the more basic we take a norm to be, the less it is an empirical question whether the agent's thought and behavior is in accord with it.

If this is so, then it does not make sense to ask, concerning a creature with propositional attitudes, whether that creature is *in general* rational, whether its attitudes and intentional actions are in accord with the basic standards of rationality. Rationality, in this primitive sense, is a condition of having thoughts at all. The question whether a creature "subscribes" to the principle of continence, or to the logic of the sentential calculus, or to the principle of total evidence for inductive reasoning, is not an empirical question. For it is only by interpreting a creature as largely in accord with these principles that we can intelligibly attribute propositional attitudes to it, or that we can raise the question whether it is in some respect irrational. We see then that my word "subscribe" is misleading. Agents can't *decide* whether or not to accept the fundamental attributes of rationality: if they are in a position to decide anything, they have those attributes. (It is no doubt for this reason that Aristotle held that an agent could not be *habitually* akratic; akrasia is deviation from a norm shared by all creatures capable of akratic acts.)

An agent cannot fail to comport most of the time with the basic norms of rationality, and it is this fact that makes irrationality possible. For if someone does fail on occasion to think or act or feel in ways that offend against those norms, he must have departed from his own standards, that is, from his usual and best modes of thought and behavior. Inner inconsistency is possible just because there are norms no agent can lack. The inconsistency does not have to be recognized by the agent, though of course it may be, nor does the existence of inconsistency depend on the agent's being able to formulate the principles against which he offends. The possibility of (objective) inconsistency depends

on nothing more than this, that an agent, a creature with propositional attitudes, must show much consistency in his thought and action, and in this sense *have* the fundamental values of rationality; yet he may depart from these, his own, norms.

To identify at least some irrationalities with inner inconsistencies, as I have in this paper, is not to explain, or even to go very far in describing, such psychological states; indeed, it makes the problems of description and explanation seem impossible. For if a person really is at a given moment harboring an inconsistent set of beliefs and attitudes, we must suppose that the views, values, and principles that create the conflict are at that moment all active tendencies or forces. It is not enough to think of one or more of the elements that create the conflict as potential and no more, or as creating a merely statistical preponderance of the rational over the irrational, where the irrational events are in the minority, but a minority expected in its numbers, and its members no more demanding explanation one by one than the events on the side of reason. Such a picture would not raise the problems here under discussion, since it would make inconsistency diachronic, not synchronic. Diachronic inconsistency is interesting in its own right, but not puzzling in the same way that synchronic inconsistency is.

Synchronic inconsistency requires that all the beliefs, desires, intentions, and principles of the agent that create the inconsistency are present at once and are in some sense in operation — are live psychic forces. It is by no means easy to conceive how a single mind can be described in this way.

We cannot, I think, ever make sense of someone's accepting a plain and obvious contradiction: no one can believe a proposition of the form (*p* and not-*p*) while appreciating that the proposition is of this form. If we attribute such a belief to someone, it is we as interpreters who have made the mistake. But if someone has inconsistent beliefs or attitudes, as I have claimed (objective) irrationality demands, then he must at times believe some proposition *p* and also believe its negation. It is between these cases that I would draw the line: someone can believe *p* and at the same time believe not-*p*; he cannot believe (*p* and not-*p*). In the possible case, of simultaneously, and in some sense actively, believing contradictory propositions, the thinker fails to put two and two (or one and one) together, even though this failure is a failure by his own (and our) standards. This is why I have urged, in several recent papers, that it is only by postulating a kind of compartmentalization of the mind that we can understand, and begin to explain, irrationality.[2]

[2] For example in Essay 2 in *Essays on Actions and Events*, "Paradoxes of Irrationality" in *Philosophical Essays on Freud*, Richard Wollheim and James Hopkins (eds.), Cambridge University Press, 1982, and "Division and Deception" in *The Divided Self*, Jon Elster (ed.), Cambridge University Press, (forthcoming).

In this paper, however, I have not attempted to describe or explain states of irrationality; I have been concerned only to show that judgments of irrationality do not have to be subjective; they may, on the contrary, be as objective as any of our attributions of thoughts, desires, and intentions.[3]

[3] An earlier draft of the present paper was discussed by John McDowell at the 1984 meeting of the Institut International de Philosophie, and I have profited from his comments.