in M. De Caro (Ed) *Interpretations and Causes. New Perspectives on Donald Davidson's Philosophy*, Synthese Library 285, Kluwer, Dordrech 1999, pp. 137-49.

Davidson on Rationality and Irrationality

Simone Gozzano

According to Davidson, the aim of the theory of interpretation is to provide the information needed for understanding the sentences of a language and, at the same time, indicate which beliefs speakers must have considering the sentences they are disposed to utter. On this point he says:

> Since we cannot hope to interpret linguistic activity without knowing what a speaker believes, and cannot found a theory of what he means as a prior discovery of his beliefs and intentions, I conclude that in interpreting utterances from scratch - in *radical* interpretation - we must somehow deliver simultaneously a theory of belief and a theory of meaning (Davidson 1974a, p.312).

The two theories to be produced simultaneously have the same goal, so that the validity of each is mirrored on the other. The theory of belief is based on the assumption that agents are rational. In this respect Davidson says: "In our need to make him make sense, we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying)" (Davidson 1970a, p. 222). So, in order to make sense of an individual, we have to suppose that most of the time, he has true and consistent beliefs. This supposition stems from the adoption of the "principle of charity". According to this principle, which Davidson uses following Quine (1960), in interpreting speakers we have to maximize the content of truth and the consistency of all their intentional states. This maximization is equivalent to the methodological rule that an interpretative mistake on behalf of the interpreter is more probable than an apparent violation of the logical principles by the interpreted individual[1]

The principle of charity is connected with a very general thesis: holism. On this Davidson says:

> Each interpretation and attribution of attitude are a move within a holistic theory, a theory necessarily governed by concern for consistency and general coherence with the truth, and it is this that sets these theories forever apart from those that describe mindless objects, or describe objects as mindless" (Davidson 1974, p. 322).

On these grounds, we may sketch Davidson's picture of rationality by saying: speakers use certain sounds as sufficient means to get certain aims given the meaning these sounds have and the beliefs the speakers suppose the hearers to have (cf. Picardi 1992a, p. 227). But this is not enough yet. Davidson thinks that we have to assume that most of the intentional states of an individual are *true*. On this he says that, even if an individual may have many erroneous beliefs:

> we can, however, take it as given that *most* beliefs are correct. The reason for this is that a belief is identified by its location in a pattern of beliefs; it is this pattern that determines the subject matter

[1] Davidson "widens" the charity principle well beyond the limits Quine imposed on it. See on this Picardi (1992a, p. 238; 1992b, p. 237) and Davidson himself (1973)]. The kind of defense that Davidson sets forth for this principle is a transcendental one: we cannot interpret speakers successfully without applying the principle. (cf. Fodor and Lepore 1992).

> of the belief, what the belief is about. Before some object in, or aspect of, the world can become part of the subject matter of a belief (true or false) there must be endless true beliefs about the subject matter. False beliefs tend to undermine the identification of the subject matter; to undermine, therefore, the validity of a description of the belief as being about that subject (Davidson 1975, p. 68).

This line of reasoning has appeal even considering the converse idea of a false belief. With respect to this Davidson says:

> To take an example, how clear are we that the ancients - some ancients -believed that the earth was flat? *This* earth? Well, this earth of ours is part of the solar system, a system partly identified by the fact that it is a gaggle of large, cool solid bodies ci rcling around a very large, hot star. If someone believes *none* of this about the earth, is it certain that it is the earth that he is thinking about? (Davidson 1975, p. 168).

So, we cannot be considered rational unless we have a large number of beliefs; moreover, these beliefs must be, for the most part, true. In this sense, the "pattern of beliefs" Davidson is talking about is an integral whole consisting of logical connections, by which we infer a belief from other beliefs. Truth goes hand in hand with consistency, and consistency is a matter of inference. So, in order to attribute beliefs, we have also to attribute inferential abilities. This point has been well expressed by Dennett also, who says: "one gets nowhere with the assumption that entity *x* has beliefs *p, q, r,* ...unless one also supposes that *x* believes what follows from *p, q, r,...*; otherwise there is no way of ruling out the prediction that *x* will, in the face of its beliefs that *p, q, r*, ...do something utterly stupid" (Dennett 1971, p. 229). From all this we may say that Davidson's view of rationality is deeply rooted in his view on truth and consistency as essential parts of the pattern of beliefs each of us should be interpreted as having.

I have insisted at some length on these points because they are crucial to the analysis of Davidson's strategy with respect to *irrationality*. There are many ways in which rationality may break down. In general, we may distinguish between akrasia and self-deception. While the first is an action contrary to the best judgment of the actor itself, the second can be conceived as the endorsement, more or less explicit, of two contradictory judgments (cf. Davidson 1970b). What I want to analyze here is the impact these phenomena have on the assumption of rationality and the specific solutions Davidson proposes in order to face it.

Both weakness of the will and self-deception are breakdowns of rationality. How can we characterize rationality in order to understand its collapsing? According to Davidson, it is a normative concept that must be accepted by the agent judged irrational. However, it is not based on a single principle, rather it is an articulated and coherent set of many requirements and principles. For instance, there is the requirement of total evidence for inductive reasoning, made clear by Carnap and Hempel, according to which we ought to act on the basis of the hypothesis best supported by the available relevant evidence (cf. Davidson 1970b, p. 41.) Also, there are the principles of logic, that ask for the consistency of the beliefs, or the principles of decision theory:

> 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 (Davidson 1985a, p. 351).

According to Davidson, a clear violation of one of these principles is a form of irrationality. But to be interpreted in this way, an agent must be interpreted as behaving against a background of rational beliefs and intentions. In a line: an act of irrationality is judged against a background of rationality. So, in order to charge of irrationality somebody

one has to contrast this judgement against the assumption of rationality. Here is the idea in Davidson's words:

> 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 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 (Ibid., p. 352).

Given this general point about the relationship between rationality and irrationality, let us see Davidson's analyses of the various cases of irrationality. The less problematic is the weakness of the will. Aristotle defined it as a form of "akrasia," that is, an action in which the actor acts against his own best judgment. The acratic is one who, knowing that in a certain situation he can do either A or B and judging that it is better to do A, nevertheless does B. In such a case the akratic is one who acts against his own best judgment, thereby violating the rational requirement of total evidence for inductive reasoning. It should be made clear, however, that the agent neither has to know nor has to explicitly endorse the requirement; acting according to the requirement is an essential feature of his being rational. However, the agent has to accept what he sees and knows as evidence for the decision, otherwise no violation of the requirement would occur.

Let me propose an example. Some time ago I was with my friend Roberto in a restaurant. Roberto told me that in Italy, when you have to choose between a bottle of sparkling water and one of natural water, it is preferable to have sparkling water. The reason for this is that the amount of mineral water produced in Italy is so elevated that one may reasonably think that, as a matter of fact, most of these waters are nothing but bottled tap water. Because adding CO2 (what makes sparkling water so sparkling) kills many bacteria, choosing sparkling water is a good way to prevent intestinal infections. While Roberto was telling me this story, I thought of a scientific article on water and public health that arrived, more or less, at the same conclusion. I agreed with Roberto's prudential advice. Moreover, in that situation, I was free to have either sparkling water or natural water. However, I choose natural water. Have I been akratic? As I sad, the problem is to understand whether I was accepting as evidence the story Roberto told me and my memory of the scientific article together with the opportunity to freely decide between the two kinds of water.

Davidson underlines that a cognitive version of akrasia is the weakness of justification. This is a case in which an agent, even having evidence to believe that *p* is more probable that *not-p*, nevertheless accepts that *not-p*. One may wonder about the relation between weakness of the will and weakness of justification. I think that the relation is a kind of entailment, in that if one acts against his own best judgment, then one should be disposed to accept the belief that supports the action or, at least, should justify the action *ad hoc*. Consider the example of the water again. If I act against Roberto's suggestion, I have to be disposed at least to justify my action in one way or another (by saying, for instance, "the bottle with natural water was closer, or something else). The motivation for asking such a justification is due to the general ability, from the agent, to provide a part of a *rationalization*, namely a reason for his actions (cf. Davidson 1963). The reason provided may be insufficient in the interpreter's eyes, but it is nevertheless what the agent may present, and usually an agent has a justification to provide (here "for no reason" would be the zero level of justification). One could argue that a further condition on weakness of will entailing weakness of justification is that the conceptual elements that determine the weakness of the will are part of our conscious experience. As we shall see, this is not so in the cases of self-deception.

Now, what is the mechanism that makes weakness of will possible? Davidson's idea hinges on the notion of rationalization. In describing and explaining an action, we look for a

primary reason, that is, a pair of state, typically a belief and a desire, that gives the reason for an agent's acting as he did. Moreover, this primary reason should be the one that has motivated the agent in his action. According to Davidson, this provides a causal explanation of the action. In cases of irrationality, however, something goes wrong. In particular, some states are *causes* but not *reasons* for whatever they cause. If we consider the example of the water again, my taking the bottle of natural water was caused by my impatience to drink water and taking the closer one was a way to speed up my thirst-quenching. Such impatience, however, is not a reason in virtue of which I may, in general, consider it preferable to drink natural water. Analogously as to justification, my saying "the bottle with natural water was closer" is not a reason to not conclude that sparkling water is, all things considered, the best option; it is only the cause of my taking natural water[2]. Our difficulty in figuring out human action is, in such cases, explained by invoking the interruption of the logical relations among mental states, and by the the presence of pure causal relations, which do not instantiate any of the rational principles and requirements mentioned earlier.

Though the idea of a cause that is not a reason for what it causes is a viable explanation for weakness of will, it is not enough with respect to self-deception. In such a case, the person who deceives himself should have a reason to endorse a weak justification and should actively consider this reason good enough to be endorsed in certain conditions (cf. Davidson 1985b).

First of all, let us try to give a definition of what self-deception is. This can be described as lying to ourselves. Let us consider, to begin with, a case of intentional deception between two agents. The condition for such a case would be, approximately, as follows.

A deceives B about *p* if and only if:

i) A knows that *p* is not the case;
ii) A acts, in presence of B, as if were the case that *p*;
iii) A avoids letting B come to know that i), instead inducing in him the belief that *p* is the case.

Now, a rough-and-ready definition of self-deception would substitute all instances of B with instances of A. From this it follows that A, at the same time, would believe that *p* and that *not-p*. Demos (1960) says that a person is self-deceived in a strong sense when he lies to himself outright, that is, when he persuades himself to believe what he knows not to be true. Here deceiving oneself involves violating the law of non-contradition.

It should be clear by now why self-deception is considered to be a *paradox* of rationality: one and the same individual in the same instant of time, holds two contradictory beliefs and does not eliminate one of them. Davidson, however, underlines that we should not consider self-deception as a phenomenon in which we hold true a single contradictory belief. An agent, he thinks, may hold that *p* and that *not-p* but cannot hold that *p and not-p* (cf. Davidson 1985a, p.353; 1985b). In this way he denies that the "Moore paradox" has any sense at all. Let us consider a case of self-deception proposed by Davidson.

Imagine an individual, whose name is Carlos, who has just bought a single lottery ticket. Carlos knows what a lottery is and how it works. Given the principle of charity, we should attribute to Carlos the belief that it is quite improbable that he will win the jackpot (let me call this the *not-p* belief). However, Carlos exhibits a series of behaviors which induce us to attribute to him the belief that it is quite probable that he will win the jackpot (the *p* belief). For instance, he incurs debts telling people that he will be very rich soon; he tries to use the

[2] Personal tastes are not, by themselves, reasons. The children who does not want the medicine because he dislike it is not offering the reason, but the cause, of his refusal.

ticket as collateral for buying goods that he could not afford in any way. (Let us suppose that he does not have any other information that we do not have, about e.g. future inheritance or the like.) In short, although Carlos knows that *p* is false (the probability factor is included in the belief content) he acts as if *p* were the case and behaves so as to reinforce his own conviction that *p* is going to be the case.

Given this example we may notice that conditions i) and iii) of the above definition are subject to two different readings. Let's consider the interpersonal definition again. According to this, A avoids letting B come to know that A knows that *p* is not the case and believes that *p* is the case. Now, what it is that A wants B not to know? A may desire that B does not discover that *p* is not the case or that A knows that *p* is not the case. That is, A may either want to induce B to think that A held a mistaken belief, or prevent him from discovering that A is a liar. Obviously, this distinction is not tenable once we substitute all instances of B with instances of A. Once one admits having deceived himself, one admits both the deceptive steps. Because admitting a self-deception is admitting both the deception and the self-directed act of deception. However, one may imagine that, in the case of self-deception, A forces himself into a form of obliviousness that *not-p* is the case, as Aristotle (*Nicomachean Ethics* 1152a 25-27.) and Pears (1985) suggest in their respective solutions. But these solutions miss that aspect of self-deception, its irrationality, and it is this which constitutes the conceptual problem it poses, that is, the acting of the agent contrary to what he knows. How is it possible to explain self-deception *qua* self-deception and, at the same time, maintain the rationality assumption as an essential aspect of the theory of interpretation? To this end Davidson considers some Freudian theses.

As we already saw, the basic mechanism that makes it possible to have cases of irrationality is to have some mental causes that are not reasons for what they cause. In the case of the lottery ticket, the supposed need for money of Carlos is the cause, but not the reason, for his endorsing the opinion that he will probably win the jackpot. This can be considered as a case of *wishful thinking*: this is a case when a desire is transformed into a belief. However, contrary to wishful thinking, in the case of self-deception the agent knows that what he believes is not supported by the available evidence. So, the mechanism of the mental causes that are not reasons is not sufficient, because the individual has to *act* in order to deceive, and acting implies an intention that, as such, has to be a reason. In this case, the desire for money would be a reason to believe that having a certain ticket makes one as the probable winner of the jackpot, but this is in clear contradiction with the other belief, about how few chances one has to win the jackpot having a single ticket. We would have, here, a "flamboyant" case of irrationality, a case in which it would be very hard to attribute any reasonable intentional state. Davidson proposes a second condition to tackle the issue of self-deception:

> Mental phenomena may cause other mental phenomena without being reasons for them, then, and still keep their character as mental, provided cause and effect are adequately segregated. The obvious and clear cases are those of social interaction. But I suggest that the idea can be applied to a single mind and person. Indeed, if we are going to explain irrationality at all, it seems we must assume that the mind can be partitioned in quasi-independent structures that interact in ways the Plato Principle cannot accept or explain. (Davidson 1982, p. 300)[3].

According to this view, then, the mind should be divided into parts, and the interaction between these parts is partially analogous to that between different minds, or persons. For instance, the fact that I yell "watch out!" pointing over your head is the cause, not the reason, of your frightened expression. Analogously, in the case of Carlos the desire for money would be in the same part, or "structure", of the mind where the belief that it is quite

[3] Plato's Principle assumes pure rationality.

improbable that he will win the jackpot, but this desire would cause the opposite belief that he will probably win, and this last belief would be in a different part of the mind. In this way the rationality of the parts would be guaranteed, because contradictory beliefs are in different parts, but there would be room for cases of irrationality, because the conditions that there are mental causes that are not reasons would be satisfied. According to Davidson, this solution has to be assumed in every theory that tries to explain irrationality, and there is no need to call for unconscious process to this end (Ibid., p. 303). In particular, Davidson thinks that "a part of the mind must show a larger degree of consistency or rationality than is attributed to the whole" (Ibid., p. 300). But the various parts are (only) partially independent, and this is the reason Davidson speaks of "overlapping territories" (Ibid., p.300, n. 6). Hence in every part, there is a structure of reasons, beliefs, intentions and attitudes connected to each other, while the interaction among the parts is due to "non-rational causality" (Ibid., p. 301).

As in the case of action theory, Davidson's model has radically changed the debate. For instance, Fingarette (1969) thought that self-deception is due to the failure of the agent to consciously recognize some aspect of his own commitments toward the world, such as decisions, activities, works, or goals. Jon Elster (1979), thought that self-deception was due to lack of information, as in the case of the Germans who pretended to ignore what was really happening to the Jews. Davidson brings our attention to the mind's structure, linking this structural question to transcendental considerations about the nature of rationality.

A criticism to Davidson's model comes from Alfred Mele. He analyses the mechanism of a mental cause that is not a reason for its effects. He thinks that this kind of mechanism can be applied also to perfectly rational cases. For instance, one may want to memorize the names of the seven hills of Rome so he can pass a test. To this end, he uses an acronym. When asked about the names of the hills, remembering the acronym is a cause, not a reason, for his emembering the names (Mele 1987, pp. 77-8). However, one might reply that Davidson is arguing that if there is irrationality then there are mental causes that are not reasons; he is not arguing that if there are mental causes that are not reasons then there is irrationality. According to Davidson, the presence of causes that are not reasons is a necessary condition for irrationality, while the presence of mental divisions is a sufficient one.

I think that the questions to be asked about Davidson's model are, instead, the following: is the model in accord with the analyses of the various interactions, causal or logical, among propositional attitudes? Is this model consistent with his general theory of interpretation and, in particular, with the holistic thesis? Let me consider, to begin with, the problem of how the parts of the mind are conceived. Let us suppose that all cases of self-deception involve minds divided into only two parts. Davidson says that these parts are in some relation. This relation, however, cannot be a logical one because, since by assumption the content of a belief is individuated by its location in a pattern of other beliefs, and given that contradictory beliefs are about the same kind of questions, we would have a logical link -- even if indirect -- between these two beliefs, and this is what the model of the divided mind is supposed to avoid. I must then exclude that the relations between different parts of the mind are rational or logical. Davidson suggests the same when he says: "The breakdown of reason-relations defines the boundary of subdivision" (Davidson 1982, p. 304). We have, then, that the links between the parts are only causal, while the rational-relations are limited to the attitudes confined within the parts. This generates a problem because, as Davidson himself notes (Ibid., pp. 301-2), if there are elements of the mind whose interaction is considered in purely causal terms without taking into account the rational description we may give of their relations, we may wonder whether, in the first instance, it is possible to explain mental facts through causal relations; and, in the second instance, what kind of generalizations we can use in explaining the functioning of the mind, since some of them are

of reason and some are of causal nature[4]. So, the hypothesis of the mental division would explain cases of irrationality at the cost of making more complex the model of the mind that lies behind the belief-desire explanation. But since the hypothesis of the mental division is based on this model, this would generate a theoretical difficulty. However, I do not intend to explore this problem further; rather I will consider whether the partitionist model is compatible with holism.

Davidson is quite aware that, at first blush, the partitioning model is not consistent with holism. He says "There is no question but that the precept of unavoidable charity in interpretation is opposed to the partitioning of the mind" (Ibid., p. 303). The reasons for this is that the adoption of the principle of charity presupposes the endorsement of holism. However, in his view this opposition is due because partitioning is invoked to allow inconstistency in the same mind while in interpretation inconsistency brings to unintellegibility. Davidson, however, resolves the problem by assuming that it is a matter of degree: are small perturbations those which generate irrationality. Large perturbations would undermine our ability to find a possible interpretation in the mental activity of any individual. I do not think this solves the problem. Arguing that there are partially autonomous mental parts with non-rational relations entails, via holism, that each single part has a complexity equivalent to the entire mind. In fact, if we want to count as a belief Carlos' belief that he will win the jackpot with the ticket he owns, we must include this mental state in a complicated pattern of other interrelated attitudes for the most part true and consistent with each other. As we saw, a mental state is a belief only in virtue of being connected in such a pattern with other attitudes. Moreover, if we want to establish our judgment of irrationality with respect to this belief, we should attribute to Carlos the opposite belief that it is quite improbable that he will win the jackpot having a single ticket, and then we have to locate this last belief in another network of attitudes that is in causal relation with the first one. Hence, in order to judge Carlos irrational, we have to double the holistic pattern of his propositional attitudes. I want to expand this point.

Let us go back to the irrational belief that is quite probable that the ticket is the right one. The general difficulty just mentioned presents itself in particular with respect to consistency. To believe that he will win the jackpot, Carlos has to have some more or less consistent beliefs about lottery, games, chance, and so on. At the same time, he should have analogous beliefs connected with the opposite belief that it is quite improbable that he will win. At this point the problem is already evident enough. However, the problem is even serious if we consider that every network or pattern is, in itself, regulated by the same principles and requirements through which we judge a pattern of attitudes to be rational. That means, each pattern has to be governed by the, usually implicit, adherence to the requirement of total evidence, to the principles of logic and those of decision theory. Interpreted in this way, the model of the partition of the mind would have two troubles: first of all it would render each of us an epistemic schizophrenic: since it may happen to each of us to hold contradictory beliefs, and since a mental state gets its content from the pattern in which is located, we should have potential patterns of beliefs available for every irrational belief we may hold. Secondly, the partitionist model has to assume an excessive epistemic redundancy. That is, each of us must be described as governed, at least, by a double set of principles and requirements for the rationality of the attitudes pertaining to each part of the mind, and each of this set of principles should sustain the beliefs enterteined in each part. Finally, imagining that an irrational individual is governed by two sets of principles regulating two partially contradictory patterns of beliefs would undermine the very interpretative practice as Davidson sees it. It is possible to see this last point in this way. If I attribute my principles to

[4] Eva Picardi has indicated to me that the problem would be that of differentiating which mental tokens are subsumed under a generalizations and which are not. This means that one should waken the generalization through ceteris paribus clause.

Carlos then, since I do not attribute to myself two sets of principles, I would not be able to attribute two sets of principles to him too. But if I do not attribute two sets of principle to him, I cannot attribute him the two patterns of beliefs that are supposedly regulated by those principles. If I cannot attribute him the two sets, then *a fortiori* I cannot attribute him two contradictory beliefs, and without contradictory beliefs there is no irrationality. From this it would follow that no one is ever self-contradictory, so that there is no self-deception, contrary to the partitionist model. It seems, then, that the partitionist model is not consistent with holism. But if this model is not consistent with holism, then it it is not compatible with the rationality assumption either, leaving us without a "background" against which to do radical interpretation[5].

One may reply that in interpreting Davidson's papers I have not applied the principle of charity. One may say that in interpreting, it is not necessary to attribute two whole patterns of beliefs, but only the contrary belief supported by a number of other attitudes sufficient to make sense of the attribution of the opposite belief. However, how are we supposed to select which beliefs we have to invoke in order to attribute the opposite belief without making an implicit appeal to the analytic/synthetic distinction? In the second place, Davidson underlines that each mental part has to be more rational that the mind in its whole. This point calls for a complex network of beliefs, not a simple one. That is, if we suppose that the part of the mind in which the contrary belief is confined is simpler than the whole mind, the problem of principles and requirements presents itself again. For if Carlos believes that he will win the jackpot, even admitting that we attribute to the network of beliefs where this belief is included only attitudes about lottery, games and so forth, if this part has to be more rational than the whole mind, it should not contain any contradiction, otherwise we are trapped in a regress. But if it does not contain contradictions, then the principles and requirements that govern it have to be strictly satisfied by the beliefs present in the network. In this case, though, the principles concerning probability and decision theory would be in sharp contrast with those pertaining to the other part of the mind. We would have, as it were, a "flamboyant" contradiction between norms that we judge fundamental and this, in turn, would condemn the individual as irrational beyond any form of interpretation[6]. Seriously taken, then, the partitionist model makes agents with contradictory beliefs unintelligible.

There is, I think, a way to save the davidsonian model. It is possible to argue that irrationality can be explained by a mental division, but that the division is between one part *within* the irrational individual, a part coinciding with his entire mind, and another one external to him, and present only virtually as a result of the projection that we, as interpreters, do. In this way irrationality would not be the result of a separation inside a single individual mind, but of a separation between an individual mind and the model of that mind that stems from the interpretation of a supposedly rational interpreter or interpretative community. An immediate consequence of this would be that there is no irrationality if there is a single individual completely alone, as in a secluded island.

This revised model originates from the idea, set forth by Davidson, according to which the propositional attitudes and the principles of the agent which create the inconsistency are present and active at once in his mind as "live psychic forces" (Davidson 1985b, p. 353). The problem, however, is that if both the belief that *p* and the belief that *not-p* were causally active in the same moment directing the agent behavior with the same intensity, how could he act at all? If we accepted Davidson's idea of the causal efficacy of all propositional attitudes, it would not be possible to explain irrational action at all. For if mental parts are of the same power as to psychic force, we would be like Buridan's ass, unable to act, and if irrationality were weaker, then we would be rational. We are left with

[5] The problem would be analogous if not worse with respect to truth, the other theory composing the general theory of interpretation

[6] Here is essential to consider what Davidson exactly says with respect to "flamboyant" irrationality

the possibility that irrational forces are stronger, but in that case were would be the forces of rationality? They would be either not causally active, or causally overpowered. But in these cases, on what basis would we postulate their presence? My proposal is that they would be postulated by an external interpreter that superimposes on the reasons of the agent those that would be the reasons that the interpreter would exhibit in those conditions, so to give a sense to the irrational utterances and actions of the agent. In passing, I want to note that if we accepted the idea that both rational and irrational forces are present at once, we would never be able to discard the hypothesis that for every rational action there is, under the "causal threshold", an irrational cause that by a deviant causal chain has generated the action.

Let's consider this revised model with respect to an example. Carlos believes that he will win the jackpot and shows this belief with utterances, actions, intentions and desires. To make sense of all his behaviors we, as interpreters, attribute to Carlos all sorts of rational beliefs and desires, considering them to be present in a purely virtual and non-causal form. We credit Carlos with a normal rationality to make sense of his irrationality. In this sense, we credit Carlos with rationality in an instrumental way. The separation is between Carlos' mind, not divided, and the mind we suppose he would have in normal conditions, that is, when no other factor intervenes to make him the way he is. Two problems arise immediately: first, whether this model eliminates irrationality; second, whether it implies the denial of first person authority.

The first problem can be stated as follows: if Carlos is irrational only with respect to us, in himself he is not irrational. On this point interpreters may oscillate. One may say that surely Carlos is irrational in himself because "in the depth of his heart" he knows that is not probable at all that he will win the jackpot. However, this objection transforms the irrationality cases in those of the weakness of the will, in which a desire is transformed into a belief. On the contrary, I think we may maintain that Carlos is committing himself to the belief that is probable that he will win the jackpot and that given this belief, it is natural that he will carry through with action based on it. If he has, as it were, "chosen" to play this game, it has to carry through with it. In short, I do not think that there is a further mysterious fact, that Carlos refuses to himself but that he knows "in the depth of his heart"[7]. This point brings us to the second one.

The problem of first person authority can be so stated: if I say "I am deceiving myself" am I victim of a self-deception? Many factors are in play. In the first place, one may argue that "I deceive myself" is an epistemic version of the liar paradox, one that does not make sense at all. In the second place, one may conceive saying to oneself "I am deceiving myself" as an act of consciousness of a deception that has already happened. In this case the one may conceive saying to oneself "I am deceiving myself" as an act of consciousness of a deception that is already happened. In this case the linguistic expression would be "I have deceived myself." What is at stake is consciousness, but we move from self-deception to the consciousness of wishful thinking, because the condition iii) of the above definition does not hold. The individual may judge that he is deceiving himself only a posteriori. In short, I think that neither the "eliminativistic" objection, nor the first person authority one pose serious troubles to the revised model. Having analysed these last points, it is time to conclude.

To sum up, I think that Davidson's model of the divided mind cannot be maintained in face of some difficulties. This for two reasons: postulating mental causes that are not reasons for their effects renders the model of the mind too complicated; even worse, holism is not consistent with the idea of a mental separation, and this has implausible consequences as to the number of beliefs and the relationships between the various principles and

[7] One may try to apply this to the case of the Germans civilians during the last world-war with respect to the Olocaust. However, it could be possible to interpret this case

requirements that govern them. As an alternative proposal, I have proposed to divide the mind, as it were, only "virtually." Mental separation would occur between the individual mind and the mind the interpreters attribute to the individual. This entails that irrationality is always judged from an external point of view.

A last observation: we saw that the theory of interpretation is composed of the theory of belief and the theory of meaning, and these two theories are somewhat interdependent. Now, if the explanation of irrationality cases hinges on the theory of belief, it hinges on the theory of interpretation too. One may wonder whether this has consequences for the general theory of interpretation. I think that the acceptance of the revised model would not affect the overall theory of interpretation. The reason for this is that the assumption of rationality, or the charity principle, is left untouched by the revised model. In this way one should not revise the theory of meaning either. So, according to the revised model, an individual could not be considered irrational in her own mind. Irrationality is always a judgment, given by an interpreter by her own lights[8].

References

Aristotle, *Nicomachean Ethics*, in W. Ross (Ed) *Works of Aristotle*, Vol. 9, Oxford University Press, London 1915.

Davidson, D. (1963) "Actions, Reasons, and Causes"; now in Davidson (1980), pp. 3-19.

Davidson, D. (1970a) "Mental Events", now in D. Davidson (1980); pp. 207-224.

Davidson, D. (1970b) "How is Weakness of the Will Oossible?", now in D. Davidson (1980), pp. 21-42.

Davidson, D. (1973) "Radical Interpretation", *Dialectica*, 27, pp. 313-28.

Davidson, D. (1974) "Belief and the Basis of Meaning", *Synthese*, 27, pp. 309-323.

Davidson, D. (1974b) "On the Very Idea of Conceptual Schema", *Proceedings and Addressings of the American Philosophical Association*, 47; pp.

Davidson, D. (1975) "Thought and Talk" in *Mind and Language: Wolfson College Lectures*, Clarendon Press, Oxford; now in Davidson (1984) pp.

Davidson, D. (1980) *Essays on Action and Events*, Oxford University Press, Oxford.

Davidson, D. (1982) "Paradoxes of Irrationality" in R. Wollheim J. Hopkins (Eds) (1982), pp. 289-305.

Davidson, D. (1984) *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford.

Davidson, D. (1985a) "Incoherence and Irrationality", *Dialectica*, 39, pp. 345-354.

Davidson, D. (1985b) "Deception and Division" in J. Elster (1985) (Ed.), *The Multiple Self*, Cambridge University Press, Cambridge.

Demos, R. (1960) "Lying to Oneself", *Journal of Philosophy*, 57, pp. 588-95.

Dennett, D.C. (1971) "Intentional Systems", *Journal of Philosophy*, 68, pp. 87-106.

Elster, J. (1979) *Ulysses and the Sirens*, Cambridge University Press, Cambridge.

Fingarette, (1969) *Self-Deception*,

Fodor, J.A. Lepore, E. (1992) *Holism*, Basil Blackwell, Oxford.

Mele, A. (1987) *Irrationality*, Oxford University Press, Oxford.

Pears, D. (1982) "Motivated irrationality, cognitive dissonance and .. " in Wollheim, Hopkins (1982)

Picardi, E. (1992a) "Donald Davidson" in M. Santambrogio (a cura di) *Introduzione alla filosofia analitica del linguaggio*, Laterza, Roma-Bari, pp. 223-66.

Picardi, E. (1992b) *Linguaggio e analisi filosofica*, Pàtron, Bologna.

Quine, W.V.O. (1969) *Word and Object*, Mit Press, Cambridge Ma.

[8] I would like to thank for comments on an earlier verion of this paper Sara Bernal (who contributed to the editing also), Mario De Caro, Eva Picardi, Antonio Rainone and Philip Robbins. This paper originates from a lecture held at the Universita' di Roma Tre. I benefit of a post-doctoral grant from this institution. I express my gratitude to this institution and to prof. Rosaria Egidi for her encouragement to my projects.

Wollheim R. Hopkins J. (Eds) (1982) *Philosophical Essays on Freud*, Cambridge University Press, Cambridge