ANNA KUSSER

Rational by Shock: A Reply to Brandt

1. It is highly controversial whether basic or intrinsic preferences and desires may be qualified as “rational” (or “irrational”), or whether on the contrary this qualification does not make sense. Richard B. Brandt – as is well known – takes side for rationality.¹ Of course much depends here on the concept of rationality one relies on. Let me therefore start by quoting Brandt’s conception of rationality as he states it at the very end of his paper: “It may be that, for a change of state of mind to be ‘rational’, it is enough if it derives from careful reflection on facts or justified beliefs, and is a change which thoughtful people generally would go through as a result of such reflection.” (Sect. 8.8)

It should be noted that this is a purely procedural definition of rationality. Certainly there are other realms of cognition for which we are able to formulate a richer concept. Nevertheless I think Brandt’s definition completely adequate in the present context. By placing “reflection” within the centre of rationality it provides a basis for further discussion without precluding it.

Since I fully agree with this informal understanding of rationality, it will not be in the focus of my critical comment. Rather I will try to raise some doubts about whether the preference changes Brandt refers to are really the results of a process of reflection. These preference changes are supposed to result from a procedure called “rational criticism”. I will argue in this paper that the processes which are supposed to take place in rational criticism and to be responsible for eventual changes of preferences are completely different from processes of reflection. If this is true then it is to be doubted whether rational criticism as Brandt conceives of it is really able to do its job.

2. Basically, rational criticism of preferences consists in confronting preferences with “demonstrated or empirically confirmed beliefs and representing them to [one]self with maximal vividness (possibly involving repetition)” (Brandt’s introductory remarks). Preferences which do not survive this confrontation are irrational whereas preferences which result from it are called rationally criticized or, for short, rational.

¹ See Brandt’s paper in this volume and also Brandt (1979), chs. 5 f.

What kind of processes are possibly induced by such a confrontation? There is a well-known mechanism by which preferences change as a function of a change of beliefs. The following is a very clear example of this mechanism: A student prefers studying at University *X* to studying at University *Y*, since she believes that Professor *B* teaches there. On receiving the information that this is false and Professor *B* actually teaches at University *Y* her preference is reversed and she now prefers to study at University *Y*.

It is easy to understand why her preference changes. Her original preference was an extrinsic or derived one. It was based on the preference to study with Professor *B* and the belief that Professor *B* teaches at University *X*. When this belief proves to be false, the derived preference vanishes, since the more fundamental preference to study with Professor *B* still stands.

No doubt, preference changes produced by this kind of mechanism are quite common and will occur on a more or less large scale if the preferences of a person are confronted with maximally corrected beliefs. However, if this mechanism alone were responsible for the preference changes constitutive of rational criticism we would have to give up the idea that it could ever apply to preferences which are not derived or intrinsic. Rational criticism would inevitably be limited to extrinsic preferences or, since a preference may be partly intrinsic and partly extrinsic, to those parts of preferences which are extrinsic. Operating on the basis of more or less articulate basic preferences, rational criticism could never change those underlying preferences themselves.

3. Clearly, Brandt wants to go farther than this (see sect. 1 of his paper).² Despite the fact that this is, therefore, not the sort of preference changing criticism he is after, I should like to make some observations regarding the mechanism at work here.

Firstly, the resulting preferences are predictable as a function of the content of the changed beliefs. This is due to a structural law which covers the connection between beliefs and derived preferences, the law being, of course, the principle of expected utility. Secondly, the newly acquired beliefs form,

² He seems to think that this sort of rational criticism holds good only for preferences for plans of action. From this one might gain the impression that whenever preferences other than for plans of action change as a consequence of a change of belief this must be due to some other mechanism. Furthermore, this mechanism might be thought to pertain to intrinsic preferences. But neither is true. In the example given above, the preference for studying at University *X* is a preference not for a plan of action but for a state of affairs. Nevertheless the change of belief immediately pertains to it. This is so because the preference is based on some other and more basic preference together with the belief in question. Hence, the crucial differentiation is not the one between preferences for plans of action and other preferences, but that between extrinsic and intrinsic preferences.

together with the more basic preferences, the sufficient cause for the preference to change. They do the causal work alone, no other causes need to be present for the change to occur. Rather, it is presupposed that no other causal influences interfere.

The first point means that the process follows a rule which secures that the change is (to quote Brandt) “generally true of thoughtful people” (sect. 8.9). The second point entails that the change of preference is not the arbitrary by-product of some mental process but precisely the result of a certain change of belief. The two points together make clear that the change of preference is determined by the content of the involved mental states, which is the reason why we may call the process in question a process of reflection. I therefore conclude that for a change of preference to be the result of a process of reflection it has to be the function of the content of the mental states which initiate it.

Concerning the process taking place within rational criticism, considerations so far have reached two results: Firstly, there must be a mechanism different from that which pertains to extrinsic preferences in the well-known way, otherwise intrinsic preferences could never be changed by rational criticism. Secondly, this mechanism must depend on the content of the involved beliefs, otherwise the whole process wouldn't be a process of reflection.

Armed with these insights I turn now to some of the examples given by Brandt to look how rational criticism is supposed to work.

4. Let me begin with three examples concerning preferences acquired by mistaken generalisation: the poor boy who acquired an aversion to furs by being traumatically frightened when playing with a rabbit (sect. 8.4), the rich man who is averse to spending money because of a destitute childhood (sect. 8.7), and the man who behaves aggressively towards his employer, whom he somehow identifies emotionally with his father (sect. 8.6). According to Brandt, in these cases the relevant preferences would vanish if confronted with a rational belief to the effect that earlier experiences were either untypical or would not occur again due to a changed situation.

But will the boy really lose his aversion if he is assured that touching furs is harmless? Will the rich man's preference not to buy luxuries really disappear if he only comes to realize that he will never have to starve again even if he spends money on superfluous things – a fact by the way which will hardly be news to him? Likewise it seems at least doubtable whether the employee would lose his aggressive attitude if only he came to understand that his employer was not his father and not even in a position comparable to his father's.

Now I admit that it is easy to cure these doubts by interpreting the above examples in such a way that a change of relevant preferences may with cer-

tainty be expected. If the boy's strong preference against touching furs is actually based on his belief that any such contact will be followed immediately by a frightening event, then of course his aversion will diminish as soon as he comes to believe that it is very improbable that a frightening event will follow.

Something parallel holds for the rich man: if his preference against buying superfluous things does actually rest on the belief that afterwards he will be badly in want of money – then, of course, his preference will immediately change if he loses this belief.

Although it is very natural to understand the examples in the way just described, this interpretation is not helpful in the present context. They are meant as examples for a change of basic or intrinsic preferences whereas here they are clear cases of extrinsic preferences the modification of which is due to the well-known mechanism explained above.

5. But how are we to understand the examples if the involved preferences are intrinsic ones? The relevant preferences don't actually have to be based on beliefs about the consequences of the preferred propositions. Such beliefs may well have played a crucial role in the genesis of those preferences, but in the course of time the preferences have gained a standing of their own. Take the example of the now rich man. He has acquired his preference against purchasing superfluous things as a consequence of the belief that he will have to starve later. But by now this preference has become a trait of his quite independent of his beliefs about his financial situation.

Given that we understand the relevant preferences in this way, it becomes again rather questionable whether they will change if there is a change of beliefs. Should a change of preferences occur, on the other hand, that would raise the strong suspicion that they had been derived ones, resting on beliefs about consequences although perhaps those beliefs were up to then quite inarticulate.

If we were not sure whether the boy's aversion to furs was intrinsic or extrinsic, we would certainly take it as a strong evidence for the first case if his aversion remained, even after he becomes convinced that furs are normally not followed by frightening events. If on the other hand his aversion did not survive this information, this would seem to be evidence that it was dependent on a more fundamental aversion to frightening events.

Prima facie there seem to be only two, mutually exclusive, cases, the one being that the relevant preferences are derived and change due to newly acquired beliefs; the other being that the relevant preferences are intrinsic and do not change if confronted with rationalized beliefs.

What Brandt tries to establish is a third case in between: that of prefer-

ences which are not derived, but do nevertheless change if confronted with relevant beliefs. In this he can only be successful if he manages to cope with two problems: Firstly, there must be a definition of intrinsic preferences which does not imply that they resist any changes of belief. Secondly it has to be explained what kinds of mechanisms are responsible for a change of intrinsic preferences as a consequence of changed beliefs.

The first problem seems to be dealt with in section 1 of Brandt's paper. Although I do not fully agree with Brandt on this point I will grant it.³ Concerning the second point his idea is that most intrinsic preferences are acquired by conditioning processes. Therefore he thinks the theory of classical conditioning the appropriate framework to make a case for modification of intrinsic preferences (sect. 6).

6. The mechanisms which he sees at work here are mainly conditioning, counterconditioning and extinction by inhibition (sects. 8.4 f.; Brandt 1979, ch. 5). It is not easy to get a clear picture of how these mechanisms do the causal work they are supposed to do. However Brandt's example in section 9.5 seems to be a good starting point to look for more details.

A man has fallen in love. His love is unreturned yet he yearns for the affection of the beloved. This is a strong preference for a state of affairs (namely to be loved by a certain person) and, let us assume, it is an intrinsic preference.

Now our lover, Brandt thinks, would lose his preference if he underwent the following treatment: whenever he is thinking of the beloved, his "reflection might be interrupted by a severe shock to the arm" (sect. 8.5). After a few such shocks he would have given up his yearning. That does not sound implausible; however, as Brandt himself observes, the change of preference would not be the result of rational criticism. Why not? Presumably because it would be the product of plain and simple counterconditioning, a process lacking the cognitive element.

Nevertheless rational criticism is supposed to function in basically the same way, the difference being only that it uses "cognitive shocks", i.e. the role of the electric shocks is assumed by beliefs: whenever the lover is musing favourably on the beloved, he has to think about displeasing attributes or

³ Brandt himself does not work with the distinction between intrinsic and extrinsic preferences but with that between preferences for plans of action and preferences *ex post* and *ex ante* for events. As already stated in the previous footnote, I think the former distinction is the one systematically required here. Intrinsic preferences are best defined as preferences which are not extrinsic, where extrinsic means derived according to the principle of expected utility. That allows for intrinsic preferences depending on beliefs, as long as it is not in the well-known way which would render them extrinsic. See Kusser (1989), part 3, III, 7.

events involving her. This would give him the purely mental or if you like “cognitive” shock that the beloved has negative attributes. Again after some effort his preferences are supposed to be gone. And if the thoughts involved here are true or rationally confirmed the whole process will constitute a piece of rational criticism.

7. But would such a change of preference really be due to a process of “reflection” and could we consequently qualify its results as “rational”? The answer is no. Having a certain belief can as such not substitute for the experience of an electric shock, since the role of the electric shock can only be taken by some other unpleasurable experience. Therefore the causal work is not done by the belief that the beloved has unpleasurable attributes, but by the fact that having such a thought constitutes an unpleasurable experience for the thinker.

A concomitant emotional experience of having a belief, not the belief itself identified by its content, is responsible for the modification of a preference. Admittedly, there seems to be a certain connection between these two, consisting in the hypothesis that to have a belief about something unpleasurable (or pleasurable) is itself an unpleasurable (or pleasurable) experience. But this tie is weak and arbitrary because we often have beliefs about unpleasurable or pleasurable things without having an emotional experience consonant with those beliefs.⁴ Moreover, according to the theory of classical conditioning there are at least two other crucial variables which determine the outcome of conditioning processes, namely temporal contiguity and frequency of conditioned stimuli.⁵ Referring to our example, the pure fact, realized by the lover, that the beloved has unpleasant attributes does not have as such a causal impact on his preferences. To ruin his yearning it needs repeated, exclusive and emotionally heavily coloured insistence on the bad attributes shortly after favourably thinking of the beloved.

Again, as in the case of the electric shock, I do not doubt that such a strategy could sooner or later cure our lover. And it is also clear that the mechanism at work here is a proper mechanism as described by the theory of classical conditioning.

Only it seems to be the wrong mechanism for the intended purposes, i.e. for rendering intrinsic preferences rational. The main variables, reflecting the causes at work here, are the emotional valence of a belief, its mode of representation, its frequency of occurrence and how it is placed within the flux of inner

⁴ Perhaps Brandt's condition that the beliefs have to be “vivid” is meant to secure that there be a consonant experience of such beliefs.

⁵ Mook (1987), chs. 7 f.

life. Even if we allow only demonstrated or empirically confirmed beliefs to affect preferences by means of those mechanisms it is never by virtue of their content that they unfold efficacy. Preference changes which follow the theory of classical conditioning are determined by features which are not systematically tied to the content of beliefs. It is therefore misleading to label preference changes which are caused in this way as consequences of “reflection on facts”.

8. The question is now whether the upshot of the previous considerations can be generalized, i.e. whether it bears on the other examples given by Brandt. I must confess that I do not always have a comparably clear picture about why preference changes should occur in the other cases, especially in the first three cases. Nevertheless the result of my argument bears fully on them as long as it is supposed that the theory of classical conditioning is the proper framework to describe them. This seems to be the case.

I conclude therefore that Brandt’s theory of change of intrinsic preferences as a consequence of confrontation with beliefs is not an adequate background for a concept of rational criticism of preferences. Even if, as Brandt claims, such preference changes did occur in the examples cited by him, that would never allow us to call the former preferences irrational.

Does that mean that the cases discussed by Brandt are all cases of rational or at least not irrational preferences? No, it only means that Brandt’s concept of rational criticism does not provide us with means to do them justice. In fact I think that Brandt really presents an impressive series of “cases in which desires manifestly are mistaken” (sect. 8).

In the remaining part of the paper I will speculate on the sources from which the plausibility of Brandt’s concept of rational preferences springs. They are not identical with the sources Brandt manifestly draws on. Therefore I shall in some degree depart from the grounds Brandt provides us with.

But as a starting point let me once again rely on one of his examples. It is about a woman who is unable to enjoy herself on Sundays, say by boating (sect. 8.9). She has acquired this aversion because she has been taught to believe that there is a loving god who wants Sundays to be kept holy. There are two ways to understand this situation: either (i) the woman acquired her preference as a consequence of this belief, but by now the preference stands on its own and is thus intrinsic or (ii) the preference is at present based on this belief and is extrinsic.

In the second case it is quite clear that the preference will change as soon as the belief is dropped. Dropping it entails that the woman no longer has to expect unpreferred consequences – say displeasing the loving god – by sailing

on Sundays. This is of course a clear case of change of a derived preference as described above. I think it is not necessary to go further into detail, let me simply remark that here pure dropping of the belief at once alters the preference.

In the first case, on the other hand, where the preference does not actually depend on the belief, the dropping of the belief as such will not affect the preference. But there is another aspect to which Brandt draws our attention, namely that the preference in question is one "preventing happiness and general desire-satisfaction" (sect. 8.9). Our lady would be better off without this preference since it prevents her from enjoying her free time. Obviously the suggestion is that if she only realized that, her aversion would disappear. This would mean that the preference is irrational in so far as it cannot confront the facts. Thinking along these lines is very obvious and seems to be the source from which the plausibility of Brandt's concept of rational criticism springs. In all of his examples the impression is conveyed that the preference in question is a serious obstacle to the person's happiness.

Take again the example of the man in love: It is of course not at all plainly mistaken to yearn for someone's affection. It is only the extra information that the affection is not reciprocated which allows us to understand that the desire is in a way mistaken, because we thereby learn that the yearning will only be a source of painful frustration and never the source of happiness. Or take the example of the man who feels bad buying luxuries. What is the problem? He prefers to leave luxuries alone thereby avoiding the unpleasant situation of spending his money on such things. But here again there seems to be an extra idea, namely that his aversion prevents him from shopping with pleasure and enjoying his purchases afterwards. Again, having other preferences would give him a better chance to be happy.

So my first thesis is that the common feature of all the plainly mistaken preferences cited by Brandt is their being obstacles to happiness and that this feature is responsible for our intuition that they are faulty. Such preferences I will call "unhappy preferences".

My second thesis concerns the suggestion that these preferences, even if intrinsic, would change if confronted with all available facts. The idea seems to be that if a person fully realizes that an intrinsic preference prevents her or him from being happy the preference will vanish. This does not hold in my opinion. The presumed change is construed in analogy to the well-known change of extrinsic preferences. The case of the puritanical lady is thus construed as follows: She wants to be happy (even on Sundays). It is a consequence of her aversion that she cannot enjoy herself by boating on Sundays. If she realized that, her aversion would decline.

Let me compare this account with the alternative account where the lady's preference itself is assumed to be extrinsic. Look at the belief involved in each case: In the case of the preference being extrinsic the belief concerns the consequences of sailing on Sundays – namely displeasing the loving god. In the case of the preference being unhappy and intrinsic the belief concerns the consequences of the aversion to enjoying oneself by boating on Sundays – namely not being able to be happy on Sundays.

The difference is this: In the first case the relevant belief concerns the consequences of *something preferred*; in the second case the belief concerns the consequences of *a person's preferring something*. The second case deals with a consequence of a psychological state of the person – namely her or his aversion. And since the consequence of this psychological state – namely not being able to enjoy oneself on Sundays – is preferred negatively, the state itself also gets preferred negatively. This is a preference about a preference, that is a second-order-preference. It is generated by a first-order-preference – the preference to be happy even on Sundays – and the belief that a certain preference is unhappy. This belief is meant to imply that there is some other preference concerning activities on Sundays which would give the woman a better chance of being happy than the preference she happens to have.

It is a further and in my eyes highly questionable step to assume that the second-order-preference has an immediate impact in that it reshapes the unhappy preference accordingly. Only if this assumption held would it be true that the belief that a preference is unhappy alters the preference in question by generating a second-order-preference which in turn changes the original unhappy first-order-preference.

In the second part of this paper I have dealt with the sources from which the plausibility of Brandt's theory originates. It was already clear from the first part that the plausibility of Brandt's concept of rational criticism lies not chiefly in his theory of change of intrinsic preferences. Rather it is tied to the series of plainly mistaken preferences he presents. The persuasive power of those examples seems to be constituted by (i) their common feature of being unhappy preferences and (ii) the suggestion that they would change if confronted with this fact.

The first point is right and needs some further clarification. The main question is, what positive concept of rationality is involved, if preferences which are obstacles to someone's happiness are accepted as mistaken? This is a call for a more precise definition of a preference being unhappy or happy.⁶

---

⁶ We have tried to work towards such a definition in Kusser/Spohn (1992), esp. part IV.

The second point seems to be false. A belief to the effect that a certain preference is unhappy does not by itself alter it. It only generates a second-order-preference. How to implement this second-order-preference by shaping the first-order-preference accordingly is an open question. Processes of conditioning and counterconditioning might be helpful here. But contrary to what Brandt thinks, they are as such not capable of rendering preferences rational. One would have to know in advance that a certain preference is irrational and how it is to be changed to qualify as rational.

References

- BRANDT (1979). Richard B. Brandt: *A Theory of the Good and the Right*, Oxford 1979.
- MOOK (1987). Douglas C. Mook: *Motivation: The Organisation of Action*, New York 1987.
- KUSSER (1989). Anna Kusser: *Dimensionen der Kritik von Wünschen*, Frankfurt/Main 1989.
- KUSSER/SPOHN (1992). Anna Kusser and Wolfgang Spohn: "The Utility of Pleasure Is a Pain for Decision Theory", *Journal of Philosophy* 89 (1992).