THE JOURNAL OF PHILOSOPHY

VOLUME LXXXIX, NO. 2, FEBRUARY 1992

SOUR GRAPES AND CHARACTER PLANNING*

> "Fit yourself into accord with the things in which your position has been cast, and love the men among whom your lot has fallen, but love them truely."
>
> Marcus Aurelius

AS a freshman in college I was first exposed to the Stoic *Lebensregel* that one should mold one's desires such that they come to match whatever it is that life has in store. Aside from being distrustful, I remember being puzzled by this ancient advice. What is puzzling about it is that, though the advice is well-taken in certain cases, it is strictly off the mark in others. Imagine a person who switched her preference for pepsi or coke back and forth depending on what the nearest vending machine had to offer. Similarly, Nietzsche's Zarathustra "laughed at the weaklings who thought themselves good because they had no claws."[1] Yet, at the same time, there is something very respectable about the capacity to exercise control over one's own emotions and desires. It was argued by Harry Frankfurt[2] that freedom of the will precisely consists in being able to control what desires will issue in action. And what

* This paper is a shortened version of the first chapter of my Ph.D. dissertation, *Reasons for Preferences,* University of Minnesota/Minneapolis. I also thank the National Fund for Scientific Research (Belgium) and the National Science Foundation (U.S.A.) for their financial support. I am grateful to the numerous individuals who have made contributions to this paper: C. A. Anderson, George Bealer, Willem deVries, Kevin Falvey, Ronald Giere, Jean Ladrière, Andrew Mason, H.E. Mason, Michael Otsuka, Michael Root, Naomi Scheman, and Yutaka Yamamoto.

[1] *Thus Spoke Zarathustra,* in *The Portable Nietzsche,* W. Kaufmann, ed. (New York: Viking, 1954), p. 230.

[2] "Freedom of the Will and the Concept of a Person," this JOURNAL, LXVIII, 1 (January 14, 1971): 5–20.

0022-362X/92/8902/57–78

could be wrong with choosing one's desires so as to avoid disappointment and frustration?

This tension in the Stoic advice is captured very crisply in Jon Elster's[3] discussion of sour grapes and character planning. These phenomena both involve a shift of desires in response to what one takes to be feasible or morally correct. For less respectable cases of this kind, the term *sour grapes* (SG) comes to mind. For more respectable cases, Elster reserves the term *character planning* (CP). He goes on to ask what distinguishes SG from CP. It is this question which articulates my earliest apprehension over the Stoic advice. Elster appeals to the autonomous nature of CP that is lacking in SG. I shall question this stand and defend an alternative solution that runs parallel to Donald Davidson's account of *akrasia*.

I

Elster extensively discusses the question what preferences must be like in order for them to be rational. Preferences acquired through SG are a paradigm case of irrational preferences. The fox in La Fontaine's fable has an appetite for grapes, though when he finds out that he cannot reach for them, he claims that grapes are too sour for his canine taste anyway. (On this reading of the fable, the fox does not change his *beliefs* about the sourness of the particular grapes in question, but rather changes his *tastes* concerning the sourness of grapes in general.[4]) Elster contrasts preference acquisition through SG with preference acquisition through CP. It is a commonplace in Aristotelian ethics that an action cannot qualify as a virtuous action unless the agent performs it with a sense of joy. This sense of joy is acquired through habituation. Imagine a poker player who enjoys the game tremendously because of the opportunities for cheating involved. She comes to realize that cheating is morally reprehensible and decides to start a better life on this score. Initially she does not find fair-play poker terribly exciting. Nonetheless, in her quest for Aristotelian virtue, she is committed to becoming the kind of person who enjoys fair play. And in order to carry through this project of CP, she chooses the Aristotelian route of habituation. That is, she decides not to give in to her initial misgivings and to stick to fair play, hoping that some day she will thus come to enjoy this style of playing.

This Aristotelian scheme can be extended beyond the moral

[3] *Sour Grapes* (New York: Cambridge, 1983), pp. 20–6, 117–9; *Explaining Technical Change* (New York: Cambridge, 1983), pp. 84–8; *Rational Choice*, Elster, ed. (New York: Blackwell, 1986), pp. 14–5.

[4] Cf. *Sour Grapes*, p. 123n.

realm. The poker player's choices were constrained by moral considerations. But choices may also be constrained simply because alternatives are no longer feasible. Imagine that the management of the casino decided to install proper lighting, such that the opportunities for cheating are entirely blocked. Our poker player is not at all pleased with this new arrangement. Nonetheless, she decides to stick with it and to try to develop a liking for fair play. It may well turn out that, after a few games, she will indeed come to enjoy fair-play poker quite a bit better than her former style.

What is puzzling about SG and CP is that both phenomena can be described in the same way, namely, both the fox and the poker player adjust their preferences in reference to what they believe to be feasible alternatives. Nonetheless, while the fox is an epitome of irrational preference acquisition, the same judgment could hardly be passed on to the poker player. Nothing stands in the way to say that her preferences took shape in an entirely rational fashion. So what is it about SG and CP that makes us pass such radically different judgments on these phenomena? The phenomena fit the same description, so what is left out of this description which can account for this distinction? Before I turn to Elster's solution to this puzzle, I shall shortcut a few incorrect, though nonetheless very telling suggestions.

(a) It may be suggested that our puzzle is ill-phrased in that it assumes there actually are two phenomena in play. One might argue that 'SG' and 'CP' describe one and the same phenomenon, which is irrational in some respects and rational in other respects. If one wants to highlight the irrational character of a particular preference adjustment in reference to the feasible alternatives, then 'SG' is the appropriate term. If one wants to highlight the rational character of the very same case, then 'CP' is the term that comes to mind. I admit that, for a great many cases of preference adjustment in reference to the feasible alternatives, it is not clear whether to categorize them as instances of SG or as instances of CP. Arguments can be made on either side of the fence and no empirical information is lacking which could decide the debate. One cannot deny, however, that there are also clear-cut cases around. It is simply incorrect to describe the fox's preference adjustment as an instance of SG. The existence of such clear-cut cases is sufficient to safeguard our puzzle. This is nonetheless a very telling suggestion. A correct account of SG and CP should also explain why there is such a broad "gray" area between both phenomena, i.e., why it is that so many cases seem to have something of SG as well as something of CP about them.

(b) It may be suggested that, whereas the poker player takes her time to come to appreciate fair play, the fox renounces his appetite for grapes at the spur of the moment. But how could this difference in timing account for our judgments concerning the irrational character of SG and the rational character of CP? Furthermore, we can well imagine a fox who passes through a phase of denial, anger, or resignation before he finally makes the SG move or a CP poker player who is very adaptable and in no time develops an appreciation for fair play. Nonetheless, this suggestion is very telling in that CP typically involves more of a project than SG. If a preference for an alternative that is no longer feasible is abandoned at the spur of the moment, there is a presumption of irrationality. This is not the case if such involves a gradual process. A correct account of SG and CP should also provide an insight into what underlies this presumption.

(c) It may be suggested that, whereas the poker player actually did change her preference, the fox merely engages in an act of self-deception. The phenomenon of SG does not consist in a preference adjustment, but rather in constructing a false belief about a stable preference for a nonfeasible alternative. I admit that there are genuine cases of self-deception in which a person forms a false belief about her preference for a nonfeasible alternative. In such cases, one will come to abandon this false belief upon sincere introspection or would indeed choose for the alternative if it were suddenly made available. Yet I want to claim that there are cases in which a person actually does change her preference because some alternative is no longer feasible and which do not qualify as instances of CP. Suppose that, upon sincere introspection, one would still hold on to one's adjusted preference or, if the nonfeasible alternative suddenly became feasible once again, one would turn away from it. I want 'SG' to refer to such cases of irrational preference acquisition, not to cases in which a false belief is being formed about a stable preference. Upon sincere introspection, the fox would still claim to have lost his appetite for the grapes and, if we were to lower the vine for him, he would turn away his head. Nonetheless, there is something very telling about this suggestion, in that, for each case of SG, there is a persistent temptation to give a description in terms of self-deception about a fixed preference. A correct account of SG should be able to explain this temptation.

II

In trying to solve the puzzle, Elster argues that preferences are rational only if they are rationally acquired and the criterion for

rational preference acquisition is autonomy. This criterion is meant to distinguish SG from CP. Unlike the fox's whims, the poker player's preference reversal in favor of fair play is autonomous in character. What is distinctive about autonomous preference acquisition is that it is open to intentional explanation.[5] Elster distinguishes between three types of scientific explanation, namely, causal, functional, and intentional. He appeals to Davidson's theory of action to fill in his account of intentional explanation.

Some of our *doings*, i.e., some of the things we do, are *actions*. Actions are doings that are intentional under some description. Other doings, like tripping over the carpet, are not intentional under any description. These are not actions, they are *mere doings*. Davidson argues that

> (I) A doing (and hence an action) is intentional under some description *d*, if and only if
>
> (a) there is some subset of the agent's mental states which provides for a reason for the doing under *d*,
>
> (b) these mental states were causally efficacious in bringing about the doing,
>
> (c) and this causal efficacity is of the correct type.

The set of mental states, which provides for a reason for an action under some description, contains a cognitive component (a belief) and an evaluative component or a *pro attitude*. For instance, a reason for the action under the description 'adding sage to the stew' is the agent's pro attitude or desire to improve the taste of the stew together with the belief that adding sage to the stew will improve its taste.[6]

Let us return to Elster's theory. A doing is open to intentional explanation if and only if there exists a description under which the doing is intentional. Preference acquisition is a doing for both the fox and the poker player. Here is how the distinction comes in between SG and CP. The poker player's preference acquisition is open to intentional explanation, i.e., there exists a description of her preference acquisition under which it is intentional. She intentionally adjusts her preferences in reference to the feasible alternatives. The fox's preference acquisition is closed to intentional explanation. He does not intentionally adjust his preferences in reference

[5] *Explaining Technical Change*, pp. 84–8.

[6] *Actions and Events* (New York: Oxford, 1980), pp. 4–12, 78–9, 86–7. Notice that Davidson himself does not use the expressions 'a doing' and 'a mere doing'.

to the feasible alternatives, nor does there exist any other description under which his doing is intentional. Consequently, the poker player's preference acquisition is an action, while the fox's preference acquisition is a mere doing. We can thus generalize: preference acquisition through CP is an action, while preference acquisition through SG is a mere doing.

What accounts for this distinction? Elster argues that the former phenomenon is caused by a metadesire to adjust one's preferences, while the latter is caused by an "affective drive"[7] that aims at reducing the frustration from unsatisfiable preferences. Metadesires, unlike drives, are pro attitudes and thus may qualify as an evaluative component in the set of mental states which provides a reason for an action under some description. Drives, on the other hand, can be no more than causes of mere doings. But what is so different about drives and desires which could motivate this distinction?

Elster believes that drives differ from desires in that they "are not conscious and known to the person who has them."[8] I have first-person knowledge of my desires as the reasons for my actions. I do not have first-person knowledge of my drives as the causes of my doings. Secondly, unlike drives, desires "may forgo short-term pleasure to achieve some longer-term gain."[9] Prudence can be an attribute only of desires, not of drives. And thirdly, Elster more than once takes recourse to mechanistic metaphors to refer to drives. According to Davidson, common mental events are in principle open to both a full physical description and a full mental description (*op. cit.*, pp. 245–9). Elster believes that drives are open only to a full physical description. Neuroscience does not yet allow for such description, however, and hence for the individuation of drives we must take recourse to their behavioral effects. If we want to describe the causal agents of these behavioral effects as such we can only invoke metaphors, like "forces" or "the wirings of the pleasure machine."[10]

Elster admits that his characterization of the set of autonomous preferences is merely tentative. First, he wants to exclude prefer-

[7] *Sour Grapes,* p. 24.
[8] *Ibid.* Cf. *ibid.,* p. 21; *Rational Choice,* p. 15; *Explaining Technical Change,* p. 87.
[9] *Sour Grapes,* p. 25.
[10] *Sour Grapes,* p. 25; *Explaining Technical Change,* pp. 71–2, with reference to A. Tversky, "Self-deception and Self-perception: Some Psychological Observations," paper presented to a colloquium on "The Multiple Self" (Paris: Maison des Sciences de l'Homme, 1982).

ences shaped by metadesires that are themselves nonautonomous in character. Second, he acknowledges that some preferences are fully autonomous, though they were not intentionally acquired. He mentions socialization within a particular culture and learning to appreciate alternatives by simply trying them out as nonintentional processes of preference acquisition. One could hardly deny that at least some preferences so acquired are autonomous in character.[11]

III

The core of Elster's theory is that preferences are rational only if they are rationally acquired and that preferences are rationally acquired if and only if they are autonomously—that is, intentionally—acquired. Since in CP preferences are intentionally acquired and in SG they are not, the former preferences are rational, the latter are not. I shall first argue that Elster's understanding of intentional preference acquisition is not in line with his professed commitment to Davidson's analysis of intentional action, nor with his own commitment to rational-choice theory. Subsequently, I shall challenge intentional preference acquisition on a strict Davidsonean reading as a criterion for rational preferences.

Consider the following application of Davidson's analysis (I) to intentional preference acquisition:

> (I^{pa}) A person intentionally acquires a preference if and only if some subset of her mental states provides for a reason for her acquiring the preference and is causally efficacious to this effect in the correct way.

Elster argues that preference acquisition through drives is not intentional, since drives are not conscious and known to the person who has them. This suggests a commitment to the claim that a person intentionally acquires a preference only if she has first-person knowledge of her reasons for doing so. Yet this claim is not in line with Davidson's analysis of intentional action. Rather, it suggests the following requirement:

> (i) An action is intentional under some description *d* only if the agent has first-person knowledge of her reasons for doing *d*.

Davidson argues convincingly against the inclusion of this condition in his analysis of intentional action.[12] Secondly, claim (i) is incompatible with two other claims to which I believe Elster is committed:

[11] *Sour Grapes*, pp. 22, 112–4, 138–9; *Explaining Technical Change*, pp. 84–7.

[12] Davidson argues that a person may intentionally poison Charles, though nonetheless be mistaken about her reasons for doing so. For instance, she may believe that her reason for doing so is to save Charles pain, while her actual reason is to have him out of the way (*op. cit.*, pp. 17–8).

(ii) Rational-choice theory can tell us more about why some person performed an action than a genuinely honest actor could (I take 'rational-choice theory' to stand for the body of theories that explain actions by means of decision-theoretic or game-theoretic models);
(iii) Explanations in rational-choice theory are paradigmatic cases of intentional explanation.

Elster is clearly committed to (iii).[13] Consider claim (ii). Many successful explanations by means of decision-theoretic or game-theoretic models unveil preferences of which the agents do not care to have first-person knowledge. The agents tell more flowery stories about the springs of their action and, what is more important, they are genuinely honest when they tell these stories. Considering the actual practice of rational-choice theory, I believe that Elster will acknowledge that this theory brings out springs of actions of which the agent does not have first-person knowledge.

Now consider these three claims. If rational-choice theory provides for a form of intentional explanation (on (iii)), it lays out the springs of actions that are intentional under some description. On (ii), these springs are at least sometimes beyond one's first-person knowledge. On (i), they must be within one's first-person knowledge. This is how the inconsistency comes in. Since Elster is committed to (ii) and (iii), he cannot hold that first-person knowledge of the reasons for one's actions is a necessary condition for intentional action. Hence, he cannot include first-person knowledge of the reasons for a preference acquisition as a necessary condition for intentional preference acquisition.

Let us now respect Elster's professed commitment to Davidson's analysis of intentional action and consider whether analysis (I^{pa}) could be a plausible criterion in Elster's theory.

First there is a challenge of principle. Preference acquisition is a type of doing. We ascribe rationality or irrationality not only to preference acquisitions, but also to doings in general. There are standard patterns in our practice of making such ascriptions to doings in general, e.g., if a doing is an action and is caused by an irrational belief (say, a belief acquired through wishful thinking), then we call the doing itself irrational. Since preference acquisitions are types of doings, a criterion for the ascription of (ir)rationality to

[13] There is Elster's explicit avowal that rational-choice explanations are a variety of intentional explanations (*Rational Choice,* p. 12). Furthermore, rational-choice theory is the focal point of Elster's discussion of intentional explanation (*Explaining Technical Change,* pp. 74–83).

preference acquisitions may be more specific than a criterion for the ascription of (ir)rationality to doings in general. The former more specific criterion cannot violate the latter more general criterion, however. A criterion for the ascription of (ir)rationality to preference acquisitions is constrained by standard patterns in our common practice of ascribing (ir)rationality to doings in general.

Let 'a description *i*' stand for a description of an action under which the action is intentional. Consider the following claims concerning the ascription of (ir)rationality to doings in general.

(i) Consider the set of all descriptions *i*. The descriptions in this set all denote actions that are intentional under *i*. Within this set, there is *both* a subset of *i*s which denote actions that are rational under *i* and a subset of *i*s which denote actions that are irrational under *i*.

(ii) Consider the set of all descriptions of doings. This set contains:
   (S_1) the subset of descriptions denoting actions that are intentional under this description;
   (S_2) the subset of descriptions denoting actions that are not intentional under this description;
   (S_3) and the subset of descriptions denoting mere doings.
   Outside (S_1) there simply is no description of a doing that is either rational or irrational under that description.

Elster clearly underwrites claim (i). Certainly an action may be rational under some description *i*. Furthermore, much of Elster's work is devoted to developing a taxonomy for actions that are irrational under some description *i*. For example, if the set of mental states which provides for a reason for an action under some description *i* contains a belief that is irrational because it is an instance of wishful thinking, then the action itself is irrational under this description.[14]

Let us turn to claim (ii). First, consider descriptions of mere doings, i.e., doings that are not actions, like tripping over the carpet. Certainly it is meaningless to say of such doings under any description that they are either rational or irrational. Secondly, consider descriptions of actions that are not themselves descriptions *i*. In 1978, Albino Luciani (John-Paul I) died less than 5 weeks after his election as pope of the Roman Catholic Church. The sentence 'The conclave elected Albino Luciani' and 'The conclave elected a pope who had less than 5 weeks to live' describe the same action, though only the former sentence is a description *i*. It may well have been rational or irrational for the conclave to elect Albino Luciani,

[14] *Sour Grapes*, pp. 26, 123–4; *Rational Choice*, pp. 13–4.

though it would be meaningless to ask whether it was rational or irrational for the conclave to elect a pope who had less than five weeks to live.

On Elster's criterion, preferences are acquired rationally if and only if they are acquired intentionally. This criterion violates the constraints on the ascription of (ir)rationality in claim (i) and (ii). On (i), if preferences are acquired intentionally, then such action may be rational or irrational under this description. On claim (ii), if preferences are acquired nonintentionally, then it is meaningless to say that such doing—regardless of whether it is an action or a mere doing—is rational or irrational under this description.

Two other challenges are pragmatical in character. Elster spells out a necessary condition on rational preferences on the basis of what he takes to be characteristic differences between SG and CP. Two questions can be asked. First, does this necessary condition—namely, intentional acquisition—indeed point to a characteristic difference between preferences acquired through SG and preferences acquired through CP? Secondly, we also hold judgments concerning the rationality of preferences outside the context of SG or CP. Does intentional acquisition indeed express a necessary condition on rational preferences in general?

Considering the former challenge, I fully agree that preference acquisition through CP is intentional in character. But why would preference acquisition through SG be any less so? Elster argues that the latter phenomenon is caused by a drive and that a drive is not a mental state and hence cannot qualify as a pro attitude in the reason for an action under some intentional description. Yet why is it that drives do not qualify as mental states? Elster mentions the absence of first-person knowledge for drives. On my earlier argument, this is a moot point, of course, since it may well be the case that a person does not have first-person knowledge of the mental states that provide for a reason for an action under some intentional description. Elster also mentions the imprudent character and the orientation toward short-term pleasure of drives.[15] Yet mental states may well share these features. My passion to cruise the Caribbean is not any less of a mental state, furnishing the pro attitude in the reason for my intentional purchase of a ticket to Barbados, when I do *not* have the necessary funds as when I *do* have the necessary funds. Hence, I

[15] Notice that this characterization is inconsistent with Elster's inclusion of the phenomenon of "forbidden fruit is sweet" as a type of preference acquisition through drives (*Sour Grapes*, pp. 111–2). Clearly such counteradaptive preferences are not geared toward (short-term) pleasure.

cannot see any good reason why preference acquisition through SG is any less intentional in character than preference acquisition through CP.

Let us now turn to the latter pragmatical challenge. Must rational preferences in general be intentionally acquired? Elster admits that there are counterexamples to his theory. There exist preferences which are indeed nonintentionally acquired, but which are autonomous and hence fully rational. He mentions preference acquisition through socialization or through learning. I take these counterexamples to be more than an indication that the theory is in need of some fine tuning.

I have restricted CP to the project of adjusting one's preferences to the feasible alternatives through the medium of habituation. Notice, however, that there are many other media for intentional preference acquisition. One such medium is precisely learning. Here is an example. I do not like dry wine, though I want to develop an appreciation for it, since all my wine connoisseurs friends serve only dry wines at their dinner parties. In order to do so, I may decide to enroll in a wine-tasting course, rather than fill up my cellar with dry wines. That is, I may choose to mold my preferences through the route of learning, rather than through the route of habituation. There is nothing irrational about preferences that are intentionally acquired through the medium of learning. But neither is there anything irrational about preferences that are nonintentionally acquired through learning. I may well come to like dry wines in a wine-tasting course, without ever having planned such preference change. This suggests that intentional acquisition may have very little to do with the ascription of rationality to preferences.

The same kind of argument can be made for other media of preference acquisition as well. CP is intentional preference acquisition through habituation. But, similarly, there is nothing irrational about nonintentional preference acquisition through habituation. After some years in North America, peanut butter has slightly moved up from the bottom of the list in my evaluation of alternative sandwich spreads. I have never planned such a thing, yet there is nothing irrational about this nonintentional preference change through habituation. In conclusion, intentional preference acquisition does not seem a promising candidate as a necessary condition on rational preferences in general.

IV

In this section, I shall venture a new solution to our puzzle. I share Elster's intuition that the phenomenon of SG yields irrational prefer-

ences while the phenomenon of CP typically[16] yields rational preferences. The challenge thus consists in constructing a necessary condition on rational preferences, such that CP can pass the test, while SG does not.

Preferences are defined over alternatives, but they are not the only kind of evaluative judgment defined over alternatives. Consider the following instructive quote from Davidson:

> . . . particular choices can be explained in decision theory along lines very similar to those followed in reason explanations: a particular action is chosen from among the available set because of the agent's beliefs (for example how apt he thinks the action is to produce various results) and the relative values he sets on the possible outcomes. His desires are thus made comparative and quantitative. Some ordinary desires, however, do not translate directly into preferences, so that not all reason explanations have a clear decision theory counterpart explanation. This is particularly evident if we think of conflicting desires. A person may have a reason for preferring A to B and another reason for preferring B to A. This is an embarassment for reason explanations, for they need to predict which reason will win out. Decision theory skips the problem, for it says nothing about why one basic outcome is preferred to another, and the theory bars possible evidence of conflict in behaviour (*op. cit.*, p. 269).

Ordinary expressions of wants, i.e., of evaluative judgments defined over alternatives, are ambiguous. If a person claims to want some alternative at some point in time, it is unclear whether she is acknowledging that she has reasons that favor the alternative—reasons that may or may not be trumped by competing reasons at that point in time—or whether she is indicating that the alternative ranks highest in her all-out ranking at that point in time. For instance, suppose I am on a strict diet and utter the words 'I really want a chocolate malt'. This may be taken to mean that I am strongly attracted to the sweet flavor of chocolate malt but may nevertheless stick to my commitment to dieting. Or it may mean that I have decided to slip on my diet and that chocolate malt has moved to the top in my all-out ranking over alternative desserts. Instances of the latter type of evaluative judgments are named "preferences" in decision theory and are represented formally by utility functions. Let us name instances of the former type of evaluative judgments *de-*

[16] I introduce this qualification because we want to allow that some cases of CP derive their irrationality from familiar patterns of irrationality, e.g., when a character planner would act on irrationally acquired beliefs.

sires. Hence, desires allow for evaluative conflict, while preferences do not. In other words, I may desire alternative *A* over alternative *B* (for some reason) *and* I may desire *B* over *A* (for some other reason), but I cannot have a preference for *A* over *B and* have a preference for *B* over *A*.[17] So what is the status of preferences within the enclosing set of evaluative judgments?

In "How is Weakness of the Will Possible?" (*op. cit.*, pp. 37–42; also 97–102) Davidson sketches an analysis of practical reasoning. In this analysis he constructs a taxonomy of evaluative judgments defined over actions. This taxonomy will be helpful in trying to understand the place of preferences within the set of evaluative judgments defined over alternatives.

Evaluative judgments defined over action tokens are either relational or nonrelational judgments. Relational judgments defined over action tokens are relative to reasons. Let *a* and *b* be two action tokens. For some reasons, I may judge that it is better to do *a* than *b*, while, for some other reasons, I may judge that it is better to do *b* than *a*. On a logical analysis, the main connective in a relational judgment is the two-place sentential operator '. . . is a ground to judge that ———', in which the first place is filled in by a set of reasons and the second place by a ranking of action tokens.

Nonrelational judgments defined over action tokens are not relative to reasons. They are the intentions that accompany our actions. If I intentionally do *a* rather than *b*, then my action is accompanied by the intention to do *a* rather than *b*. Davidson takes this intention to be the nonrelational judgment that *a* is better than *b*.

Reasons for relational judgments defined over action tokens come in pairs, each containing a relational judgment defined over action types and a belief ascribing action tokens to action types. Here is an example. On this cold winter night, I am considering whether I shall walk my dog or read a book in front of the fireplace. On the one hand, I may think it is better to walk my dog than read a book, considering my dog needs the exercise. On the other hand, I may think it is better to read a book, considering it is so cold out. On

[17] Davidson's claim that a person may have a reason for *prefer*ring *A* to *B* and another reason for *prefer*ring *B* to *A* is slightly confusing. While *preferences* in a technical sense do not allow for evaluative conflict of this kind, Davidson apparently does not carry over this technical sense from 'preference' to the related verbal form 'preferring'. I assume that 'preferring *A* to *B*' is here being used in a nontechnical sense and is not the kind of evaluative judgment that is at work in a decision theory, i.e., it is not synonymous with 'having a preference for *A* over *B*' in a decision-theoretic sense.

a logical analysis, my reason to think it is better to walk my dog than read a book is the relational judgment defined over action types:

(M_1) That some action token x is an instance of the action type 'giving proper care to pets' and some action token y is an instance of the action type 'neglecting pets' is a ground to judge that x is better than y.

and the belief:

(m_1) The action token 'my walking my dog (now)' is an instance of the action type 'taking proper care of my dog' and the action token 'my reading a book (now)' is an instance of the action type 'neglecting pets'.

The conclusion of this reason pair is the relational judgment defined over action tokens:

(C_1) (M_1) and (m_1) are a ground to judge that the action token 'my walking my dog (now)' is better than the action token 'my reading a book (now)'.

Analogously, a conflicting relational judgment defined over action tokens (C_2) can be formed which is relative to the relational judgment defined over the action types 'exposing myself to the cold weather' and 'not exposing myself to the cold weather' (M_2) and the matching belief ascribing the action tokens in question to these action types (m_2):

(C_2) (M_2) and (m_2) are a ground to judge that the action token 'my reading a book (now)' is better than the action token 'my walking my dog (now)'.

There is a particular relational judgment defined over action tokens in which we take a special interest, namely, the relational judgment that is relative to *all* of the agent's reasons, i.e., all of the mental states that the agent takes to be relevant to the choice at hand. Such relational judgment is an *all-things-considered judgment*. Singular reason pairs may lead to conflicting relational judgments defined over action tokens. The all-things-considered judgment is then determined by the relative importance that the agent assigns to each reason pair. All-things-considered judgments do not follow from the agent's reasons as a matter of logic. I myself am a pet lover and the cold weather does not throw me off that easily. So here is my brave all-things-considered judgment:

(C_{all}) M_1, m_1, M_2, and m_2 are a ground to judge that 'my walking my dog (now)' is better than 'my reading a book (now)'.

Nonetheless, I may stay home and read a book in front of the fireplace. How so? Either I may not have formed an all-things-considered judgment at all, because I did not give sufficient thought to the choice at hand. Or I may have formed such judgment but, through weakness of the will, I did not act in accordance with it. In both cases, the intention or nonrelational judgment that accompanies my action is not informed by the all-things-considered judgment, but rather by the relational judgment (C_2), which is relative to a subset (containing M_2 and m_2) of the complete set of reasons.

On Davidson's principle of continence, an action is rational only if the intention that accompanies it is informed by the agent's all-things-considered judgment for the choice at hand. Both ill-considered as well as *akratic* actions fail to pass this criterion of rational action.

What can be learned from Davidson's taxonomy for understanding the place of preferences in the set of evaluative judgments defined over alternatives? Let alternatives be *state-of-affairs tokens* (SOA tokens). For instance, my preference for this particular Volvo over that particular Austin *is* my preference for the SOA token of 'my owning this Volvo (now)' over the SOA token of 'my owning that Austin (now)'.

Preferences are a special kind of judgment defined over SOA tokens in that they do not allow for evaluative conflict. For some reasons I may desire the Volvo rather than the Austin and for some other reasons I may desire the Austin rather than the Volvo, but I cannot have a preference for the Volvo over the Austin and have a preference for the Austin over the Volvo. This should ring a bell. I propose that preferences, just like intentions, are nonrelational judgments. Other evaluative judgments defined over SOA tokens are relational judgments, i.e., they are relative to reasons. And this parallel can be carried even further. Reasons for relational judgments defined over SOA tokens come in pairs, each containing a relational judgment defined over SOA types and a matching belief ascribing SOA tokens to SOA types. For example, a reason for desiring this Volvo rather than that Austin is the relational judgment defined over SOA types:

(M_1^*) That a SOA token x is an instance of the SOA type 'owning a Swedish car' and a SOA token y is an instance of the SOA type 'owning a British car' is a ground to judge that x is better than y.

and the belief:

(m_1^*) The SOA token 'my owning this Volvo (now)' is an instance of the SOA type 'owning a Swedish car' and the SOA token 'my owning this Austin (now)' is an instance of the SOA type 'owning a British car'.

This reason pair implies the relational judgment defined over SOA tokens:

(C_1^*) (M_1) and (m_1) are a ground to judge that the SOA token 'my owning this Volvo (now)' is better than the SOA token 'my owning this Austin (now)'.

Again, different reason pairs may support conflicting relational judgments defined over SOA tokens. A particular relational judgment in which we take a special interest is the all-things-considered judgment defined over SOA tokens. This judgment is relative to *all* of the person's reasons, i.e., all of the mental states that the person takes to be relevant to the ranking at hand.

Davidson's principle of continence imposes a constraint on actions. This principle can also be understood to impose a constraint on the intentions accompanying the actions. The intention to do *a* rather than *b,* i.e., the nonrelational judgment defined over action tokens that it is better to do *a* than *b,* is rational only if it is informed by the all-things-considered judgment for the choice at hand. I propose to construct an analogous constraint for preferences:

(R) A preference for SOA token *s* over SOA token *t,* i.e., the nonrelational judgment defined over SOA tokens that *s* is better than *t,* is rational only if it is informed by the all-things-considered judgment for the ranking at hand.

All-things-considered judgments are relational judgments that are relative to all of the person's reasons, i.e., all of the pairs—each containing a relational judgment defined over SOA types and a belief ascribing SOA tokens to SOA types—which the person takes to be relevant to the ranking at hand. Let us name a relational judgment defined over SOA tokens which is relative to *one* such reason pair *a criterial judgment.* An example of such criterial judgment is (C_1^*), i.e., on a more colloquial version, my judgment that, considering its country of origin, I would rather have the Volvo than the Austin. But there may also be other criteria, which the person takes

to be relevant to her ranking, say, color, year, etc. The role of the all-things-considered judgment is to arbitrate between criterial judgments. If all criterial judgments favor the same alternative, arbitration simply means acknowledging the unanimous vote. If different criterial judgments favor different alternatives, arbitration involves an assessment of the relative importance that each consideration carries in one's mental life. I can thus rephrase the constraint on preferences:

(R′) A preference for SOA token *s* over SOA token *t* is rational only if it results from proper arbitration over *all* criterial judgments defined over *s* and *t*, implied by the reason pairs that the person takes to be relevant to the ranking at hand.[18]

How do we go about judging the propriety of such arbitration? For our purposes, it is sufficient to see that there are clear cases of preferences that are irrational precisely because they did not result from proper arbitration over all criterial judgments. For instance, suppose I get carried away by the wonderful stereo equipment in the Austin. I thus set my preference for the Austin over the Volvo, overlooking that the Austin is an older, British car in an ugly ochre. On due reflection, I must admit that the consideration that set my preference carries much less importance for me than the considerations that were overridden by it. It can be agreed that such would indeed be an irrational preference.

I shall now argue that this necessary condition on rational preferences can solve our puzzle, i.e., can drive a wedge between SG and CP. Let alternatives *s* and *t* both be feasible alternatives to begin with. I have a preference for *s* over *t* and this preference is informed by an all-things-considered judgment, i.e., it results from proper arbitration over all criterial judgments. It now turns out that, for some reason, alternative *s* is no longer feasible anymore. In response, I intentionally adjust my preference such that I come to have a preference for *t* over *s*. So far, SG and CP run entirely parallel. They also run parallel in that I may or may not have first-person knowledge of my reasons for adjusting this preference, i.e., I

[18] One may object that no more is required for rational preferences than that one properly arbitrate over some subset of the relevant criterial judgments which is commensurate with the prominence that holding preferences over the alternatives in question has in one's life. This objection is well-taken and can be accommodated for in criteria (R) and (R′) and in the subsequent discussion at a serious cost of stylistic simplicity, though without affecting the general argument.

may or may not know that I resort to this preference adjustment *because* I realize that my favorite alternative is no longer feasible anymore.

The phenomena do differ, however, in the following respect. In a typical case of SG, I adjust my preference, though I do *not* adjust any of the relational judgments defined over the SOA types, nor the relative importance I attach to each consideration. Assuming fixed beliefs, the original set of reason pairs thus remains constant and there is no change in the relative importance that each pair carries in my mental life. Since the all-things-considered judgment for the ranking at hand is relative to this complete set of reasons, it too remains fixed over the course of the preference adjustment. In other words, nothing changes about the arbitration problem over the criterial judgments for the ranking at hand. Hence, the adjusted preference cannot be informed by the all-things-considered judgment, i.e., it cannot result from proper arbitration over all criterial judgments. And this is what accounts for the irrational character of preferences acquired through SG.

A typical case of CP is the more involved project in which I adjust my preference as well as my reasons for the ranking at hand. That is, either I shift around some of the relevant relational judgments defined over SOA types (which, together with fixed beliefs, imply a change in criterial judgments), or I change the relative importance that I attach to each consideration. Consequently, relative to my adjusted set of reason pairs and the relative importance I assign to them, I now think that the still feasible alternative t is better than the no longer feasible alternative s. In my adjusted all-things-considered judgment, alternative t moves up above alternative s. Hence, over the course of the CP move, my preference in flux remains informed by the all-things-considered judgment in flux. In other words, my preference results at all times from proper arbitration over the relevant criterial judgments. And this is why there is nothing irrational about preferences acquired through CP.

Let us turn to an example. For quite a while I have been interested in a particular new job, which would give me a chance to work overtime at a decent pay and with challenging career opportunities. I then find out that this change of jobs is no longer feasible for some reason. In response, I shift around my preference such that I come to prefer my actual job to the new job.

In a typical case of SG, this preference adjustment is the only evaluative shift in my mental life. I still hold the relational judgment over SOA types that 'having a job with opportunities for well-paid

overtime work' is better than 'having a job which leaves time for leisure' and that 'having a job with challenging career opportunities' is better than 'having a more stable and less stressful job'. The set of relational judgments defined over SOA types and of beliefs ascribing SOA tokens to SOA types ('my having the new job' is an instance of 'having a job with opportunities for well-paid overtime work', etc.) still imply a pair of criterial judgments, which cast a unanimous vote in favor of the new job. I still am the kind of person who has all reason to prefer the new job to my actual job. If I do nonetheless have a preference for my actual job over the new job, then my preference is not informed by the all-things-considered judgment for the ranking at hand. In other words, my preference does not result from proper arbitration over the relevant criterial judgments, which, in this particular case, would be acknowledging their unanimous vote. Consequently, my adjusted preference in favor of my actual job can save me from frustration only at the cost of being irrational.

In a typical case of CP, I effect a more radical change in my mental life. Not only do I come to hold a preference for my actual job, but I also bring about a shift in the set of relational judgments defined over SOA types such that, indeed, I become the kind of person who genuinely has reason to hold the adjusted preference. I now appreciate the charm of a job which allows me more time for leisure and in which there is no pressure for promotion. My adjusted preference is thus informed by the all-things-considered judgment that is relative to my newly acquired relational judgments defined over SOA types together with fixed beliefs. That is, it results from proper arbitration over the newly acquired criterial judgments, which now cast a unanimous vote for my actual job. Hence, my adjusted preference in favor of my actual job passes through as a fully rational preference.

What about La Fontaine's fox? The fox loses his appetite for grapes when he finds out that he cannot reach for them. We do not assume, however, that he has changed his relational judgment defined over SOA types that, say, 'indulging in juicy summer fruits' is better than 'keeping to a more wholesome diet'. Over the course of the SG move, he remained the kind of fox who has a craving for juicy summer fruits in general. And what about the poker player? The CP poker player did not only change her preference concerning fair-play poker in the well-lit Casino, but she also came to hold the relational judgment defined over SOA types that 'playing a fair game' is better, is more fun than 'cheating'. Over the course of the

CP move, she has become the kind of person who genuinely values fair play in general.[19]

Both the unsuccessful job hunter and the poker player changed around *all* relevant relational judgments defined over SOA types in their CP moves. Notice that CP may well take a more subtle course. It may be sufficient to change a proper subset of the relevant relational judgment defined over SOA types to effect the desired shift in the ranking of the all-things-considered judgment. Or, in case of conflicting criterial judgments, it may be sufficient to adjust the relative importance attached to the respective considerations to have proper arbitration result in the desired preference shift.

V

In section I, I considered three responses to our puzzle. I argued that, though none of these responses was correct, they were all in some way very telling suggestions. A proper solution to our puzzle should provide some insight as to why it is that each suggestion does indeed seem prima facie plausible. Let us consider whether my theory of rational preferences can meet this challenge.

(a) Between clear-cut cases of SG and CP, there is a broad "gray" area of preference adjustments that are undecided as to whether SG or CP is the more appropriate classification. Can my theory account for this phenomenon? In a typical case of SG, my complete set of reasons for the ranking at hand remains untouched over the course of the preference adjustment. In a typical case of CP, I adjust both this set of reasons and my preference such that the adjusted preference is informed by the all-things-considered judgment that is relative to the adjusted set of reasons. In between these typical cases, there is a continuum of preference adjustments, in which I actually do shift around *some* relational judgments defined over SOA types or make *some* changes in the relative importance I attach to each consideration, though none of this is quite enough to let the balance in the all-things-considered judgment tip over in favor of the feasible alternative. Some steps are being taken in adjusting the set of reasons for the ranking at hand, though not quite enough for the adjusted preference to pass through as a rational preference. Between typical cases of SG and typical cases of CP there is a continuum of preference adjustments accompanied by not entirely successful shifts in the set of reasons for the ranking at hand.

[19] There are certain interesting worries involving (a) an attempt to blur the distinction between CP and SG by allowing for relational judgments that favor feasible SOA types, (b) radical evaluative shifts that are nonetheless irrational in character, and (c) shifts from irrational to rational preferences which are motivated by feasibility considerations. I intend to address these objections in a later paper.

Are preference adjustments on this continuum watered-down cases of SG or of CP? This is merely a verbal point. On the one hand, it seems appropriate to talk about SG for all cases on the continuum, because of the irrational character of the thus acquired preferences. On the other hand, if a person has made serious changes in her relational judgments defined over SOA types, or in the relative importance she attaches to the criteria involved, it seems appropriate to talk about CP, even though these changes may not be quite sufficient to support her preference adjustment. Many real-life or fictional cases of preference adjustment are actually located on this continuum. They are neither typical cases of SG, nor typical cases of CP. My theory shows how there is something to be said for classifying such cases in either category. And this accounts for the phenomenon that there is indeed a broad range of cases in which we are undecided as to whether 'SG' or 'CP' is the more appropriate term.

(b) The SG move *typically* is a response at the spur of the moment, while CP *typically* is more of a long-term process. If preferences are abandoned at the spur of the moment, there is a presumption of irrationality, which there is not if such occurs gradually. My solution to our puzzle can readily account for this presumption. While SG targets only one's preference, CP focuses on the complete chunk of one's mental life which is relevant to the preference in question. SG involves one single adjustment, CP involves a project of change. A sudden change of preference in response to considerations of feasibility thus makes us suspicious that SG rather than CP is in play.

(c) In trying to give an account of SG, there is always the temptation to call in self-deception, even so if the fox would deny his appetite for the grapes upon sincere introspection and would turn his head if we were to lower the vine for him. Can my solution account for this temptation?

In attributing a preference to a person we rely on presumptions. One such presumption is that, if a person acknowledges a preference upon sincere introspection, then she does actually hold such preference. Another presumption is that a person's choices are indicative of her actual preferences. Certainly, both presumptions may clash and, within particular circumstances, one presumption may seem to deserve more credibility than the other. For instance, if the controversial choice is made under stress, then we tend to put more trust in the introspective presumption. On the other hand, if acknowledging some preference is threatening to one's self-image, then we tend to put more trust in the behavioral presumption.

There is also a third presumption that comes into play, namely, a principle of charity in practical reasoning[20]: if a person holds a set of reasons which makes it rational for her to hold one preference rather than some other, then there is a presumption that she actually does hold this preference. In claiming that the fox actually did change his preference, this presumption of rationality is not respected. On the other hand, if we were to claim that the fox did not change his preference, then we would not respect both the introspective and the behavioral presumption.

I do not want to suggest that the number of presumptions can tip the scale. Rather, I propose to restrict the term 'SG' to those circumstances in which we tend to put our trust into the introspective and the behavioral presumption, rather than into the presumption of rationality. But even though one may feel most at ease in trusting the introspective and the behavioral presumption, the temptation still lingers to interpret the phenomenon in reference to a principle of charity in practical reasoning, i.e., to deny that any preference adjustment has taken place. If this temptation were to be followed, one would need to take recourse to self-deception to make up for not respecting the introspective presumption. Hence, my theory of rational preferences drives a wedge between SG and CP without describing the former phenomenon in terms of self-deception. At the same time, it can explain why there is a prima facie plausibility to such description.

LUC BOVENS

University of Colorado/Boulder

[20] Davidson, *op. cit.*, pp. 221–2; and "Paradoxes of Irrationality," in R. Wollheim and J. Hopkins, eds., *Philosophical Essays on Freud* (New York: Cambridge, 1982), pp. 289–305.