

Is non-instrumental valuation of processes rational?

Chapter 1 of Ph.D. dissertation: *Explorations in Process-Dependent Preference Theory*

Martin E. Sandbu

Harvard University

sandbu@post.harvard.edu

September 2003

Abstract

A recent turn in preference theory has been the consideration that preferences over outcomes may be sensitive to the process by which the outcomes are brought about. This possibility has motivated Amartya Sen's (1997, 1999) distinction between preferences over *comprehensive outcomes* and *culmination outcomes*. This essay examines the question of whether preferences which rank these differently could be rational. I present two arguments why they may be thought irrational. First, process-dependence seems like sensitivity to sunk costs, which is typically thought to be irrational. Second, I show that such preferences may be *self-defeating*, since they promote actions that undermine the source of their own value. The first of these arguments is flawed. The second needs to be rebutted, and I offer two accounts of how the charge of self-defeatingness can be avoided. The first is that directly non-instrumental valuation of processes could be indirectly instrumental. The second is that actions can derive value not just through instrumental relations to other valuable things, but also through symbolic or evidential relations. Preferences over comprehensive outcomes can therefore be rational, but which of the two reasons accounts for their rationality matters for how they should be treated in social evaluation.

1 Introduction

Individuals have values. Some of these values are about outcomes, narrowly defined — we value some states of affairs more highly than others. Others are values about processes — we prefer arriving at a state of affairs in a certain way rather than another. The relative importance of these two kinds of values differs across people. Some do not mind how things are done as long as the result is right; others are willing to sacrifice a lot of valuable outcomes rather than achieve them in a way they find unacceptable. Evaluative theories are like people in this respect. Some, like welfare economics in its conventional form, focus uniquely on end-states. Others, like Nozickian libertarianism (Nozick 1974), are pure procedural theories. Many have features of both. John Rawls’s (1971) theory of justice, for example, uses a concept of pure procedural justice, but also includes substantive criteria for justice that specify distributive outcomes.

Amartya Sen (1997, 1999) has recently highlighted this distinction in theories *about* values, in particular formal preference and choice theory. He accuses conventional economic theory of gratuitously assuming that people only have preferences over what he calls “culmination outcomes” — outcomes described simply as states of affairs, without reference to their histories — and not admitting that people may have preferences defined over “comprehensive outcomes” whose description includes the actions or processes that brought them about. It is undeniable that people often state this more complex kind of preference. There is a question, however, as to how best to understand such statements. Are they “genuine” preferences like plain preferences over culmination outcomes? Are they really a shorthand for preferences over culmination outcomes that incorporate a sophisticated calculation of complex and far-fetched consequences? Or do they simply reflect cognitive limitations, and ought to be seen as confusions, errors, or superstitions that people would abandon if they managed to think through them clearly? In particular, there has been no discussion of whether such preferences are *rational*. If they are not, there is a presumption that they may not need to be taken into account. And the divergence between the different moral theories mentioned above as to whether or not processes matter suggests that the question of whether it is rational to have preferences over processes is not satisfactorily resolved. I suspect that many

think the answer to the question obvious, but also that people do not agree about what the obviously right answer is. Utilitarian philosophers and economists often only consider preferences over culmination outcomes. Rights theorists and contractarians typically think that procedural values are separately important or even that they have a lexicographic priority over concerns about culmination outcomes. The rationality of caring about processes is not a subject on which consensus reigns.

Nor is the question of merely intellectual interest. Almost all theories of social evaluation, and probably all acceptable ones, give some role to individual preferences in the determination of how a society should be organised and governed. At the same time, it is clear that rational and irrational preferences are not on a par in how they ought to enter social evaluation. While it may be wrong to ignore irrational preferences altogether, they call for a different treatment and most likely ought to be discounted compared with rationally defensible preferences. The status of preferences over comprehensive outcomes, therefore, matters for the relative appeal of different social theories. If such preferences are rational, purely outcome-based theories like conventional welfare economics and cost-benefit analysis can be accused of bias against procedural values. If they are irrational, that poses a problem for moral or political theories that place a value on processes that is independent of their instrumental effects on outcomes. Moreover, even if preferences over comprehensive outcomes are rational, the way in which they can be shown to be rational may matter for how they enter social evaluation. Whether we deem preferences over comprehensive outcomes rational or irrational, therefore, is a prerequisite for choosing among theories of justice.

I begin the analysis by elucidating what I mean by the rationality of preferences (section 2). Section 3 discusses how preferences over comprehensive outcomes are characterised by the property of *process-dependence*. I then consider three ways in which actions may be valued. Direct instrumental valuation and intrinsic valuation of actions are unproblematic from the point of view of rationality. But direct instrumental valuation cannot account for process-dependence, whereas many cases of preferences over comprehensive outcomes cannot be seen as valuing actions intrinsically. Process-dependence could also reflect actions that are valued *derivatively but non-*

instrumentally. This is the kind of preference I focus on in most of the essay. Section 4 presents two arguments why process-dependent preferences are irrational. First, process-dependence resembles caring about sunk costs, commonly thought to be irrational. Second, when such preferences endow actions with derivative non-instrumental value, they risk being self-defeating. The first of these arguments fails. The second needs to be seriously addressed. Sections 5 and 6 offer two alternative accounts of why, when properly understood, value systems defined over comprehensive outcomes do not exhibit the kind of inconsistency which rationality must condemn. The first account is based on the *indirect* instrumental value of such preferences. The second argues that instrumentality is not the only relation through which actions can derive value. Which of these accounts is correct has consequences for whether or not preferences over processes need to be taken into account in social evaluation. I finish with an observation about the distinction between instrumentalism and consequentialism — value systems defined over comprehensive outcomes involve non-instrumental valuation but can remain consequentialist (section 7). Section 8 concludes.

2 Defining the question

What does it mean to have non-instrumental preferences for procedures, or to have preferences over comprehensive outcomes? And what would it mean for such preferences to be irrational? We shall focus here on *consistency* as a necessary condition for rationality.¹ Consider the analogous question of the rationality of beliefs. To be rational, a belief system must be consistent. An inconsistent system of beliefs cannot be sustained in the face of reasoned scrutiny. We may restate the question, therefore, as asking whether preferences over comprehensive outcomes are inconsistent in a way that rationality rules out. The answer to that question is going to depend on what we mean

¹Consistency is a necessary but not a sufficient condition for rationality. For a belief system to be rational, it must be consistent but also meet other conditions; the beliefs must for example have been generated by a rationally acceptable mechanism. Similar additional conditions hold for the rationality of preferences. (See Nozick (1993) for an ample list of possible conditions for the rationality of beliefs and preferences.) When we choose to focus on consistency, it is because it is a minimal condition. If preferences over comprehensive outcomes are to be rational, they must *at least* satisfy some consistency conditions. That is not to say that they do not also have to satisfy other conditions. However, since one can make a case that they do not even satisfy consistency, as I discuss in section 4, that case has to be rebutted first.

by “preferences.” For preferences to be under the jurisdiction of rationality at all, for there to be a meaningful question of whether a given pattern of preferences is rational or irrational, our notion of preference must be one that admits of the kind of inconsistency that rationality rejects. Since the concept of preference — and its sibling “utility” — has many different uses, some of which do not admit of such inconsistencies as fall under the judgment of rationality, it is necessary at this point to clarify the sense in which “preference” is employed in this essay.

The many uses to which the terms “preference” and “utility” have been put are a reflection of how useful the concepts are to social theorists of various disciplines. Yet this profusion of interpretations can be a source of confusion if not of serious fallacy.² At one end of the spectrum of possible meanings is preference as *well-being*. In this usage, which we may call the *eudaimonistic* sense of preference, preferring x to y means that x produces more well-being (to the person whose preference it is) than y . There are many sub-theories of eudaimonistic preference theory, in which well-being is understood for instance as pleasure, happiness, desire-satisfaction, or some other type of contentment. Depending on the theory, preferring x to y means desiring x more strongly than y , or obtaining more happiness or pleasure or satisfaction from a realisation of x than of y . We do not need to distinguish between these sub-theories for our purposes. They all share well-known problems; for example, it is not clear that all objects of preference can be measured with a single kind of desire or pleasure. Desires, pleasures and happiness come in different flavours, not all of which are commensurable. The most important question for us, however, is whether these theories even admit of any talk of the rationality of preference. The various notions of well-being are all in some sense raw feelings, which cannot be inconsistent in the way beliefs are. Inconsistent beliefs *contradict* each other as a matter of logic, which sources of well-being do not. We may of course call desires “inconsistent” if, say, they cannot be mutually satisfied and their coexistence causes frustration. But this is not a kind of inconsistency that is threatened by reasoned scrutiny. A better word is *incompatibility* of sources of well-being, and that incompatibility is a fact of the world, not a logical contradiction.³ Preference in the eudaimonistic sense just does not admit of the kind of

²See Sen (1973), from which the next few paragraphs borrow extensively.

³Psychologists also speak of “dissonance”; see (Festinger and Carlsmith 1959, Festinger 1962).

inconsistency that rationality abhors; such preferences are *arational*, beyond the domain of reason. This argument goes back to Hume's famous dictum that "reason is, and ought only to be, the slave of the passions."⁴

At the other end of the spectrum is the notion of "revealed preference," popular in economics and first formulated by Paul Samuelson (1938). In this interpretation "preference" is just the induced ranking of alternatives that can be derived from observed choice behaviour. Preferring *x* to *y*, on the revealed preference view, simply means "choosing *x* when *y* is available," or something of that nature. Rationality may seem to have more bite on this type of preference. The theory of revealed preference says that we can determine whether preferences are rational by checking whether choices conform to certain conditions of *internal consistency of choice*, dubbed "axioms of revealed preference." Indeed this is the prevailing definition of "rationality" in economics (see for example Andreoni and Miller (2002) for an investigation into the "rationality" of altruistic choices). Amartya Sen (1993, 1995) has convincingly argued against this position, however, by showing that there can be no coherent notion of *purely* internal consistency of choice. How can we tell from looking purely at a set of choices whether they are internally inconsistent? A decision-maker is not necessarily inconsistent if she chooses, say, chocolate ice cream over strawberry ice cream today, but makes exactly the opposite choice tomorrow. She may consider the options differently depending on when the choice has to be made. Whenever we face a set of choices that is alleged to be internally inconsistent, we can always redescribe and individuate the alternatives so as to make them internally consistent. In order to claim that there is something inconsistent or contradictory about the choices, we need to know something about the agent's *non-revealed*

⁴Derek Parfit argues in *Reasons and Persons* that desires can be irrational: "It is irrational to desire something that is in no respect worth having, or is worth avoiding." However, he admits that "there may be few actual desires that are irrational in this way," and goes on to discussing *second-order* desires. "It is at this secondary level that the charge 'irrational' can be most plausibly made. Someone is not irrational simply because he finds one experience more painful than another. But he may be irrational if, when he has to undergo one of these two experiences, he prefers the one that will be more painful" (Parfit 1984, p. 123). These second-order "desires," I submit, are closer to what I later in this section call "values," which are indeed subject to reasoned scrutiny. In his example of "future-Tuesday-indifference," a man who cares strongly about avoiding future pains unless they happen on a Tuesday, for no reason, Parfit does not say that the man has an irrational *desire* for pain on future Tuesdays. He instead uses expressions like "[t]his man's *pattern of concern* is irrational," "[p]referring the worse of two pains, for no reason, is irrational" (Parfit 1984, p. 124, my italics). Below I discuss future-Tuesday-indifference as applied to values more at length.

preferences, so that we may correctly describe the choice set. Even to talk about the rationality of choice behaviour, we cannot use a concept of preference that is completely behavioural and eschews all references to mental constructs. While the well-being notions of preference admitted of a kind of inconsistency but not one that falls within the jurisdiction of rationality, here we have the opposite case. Even if rationality may have something to say about the consistency of behaviour, *mere* choices cannot fail to be consistent, under the appropriate redescription. To give rationality something to scrutinise, we have to go beyond *internal* consistency of choice towards a theory of what the choices ought rationally to be consistent *with*.

Neither of these two preference concepts,⁵ therefore, will serve our purpose. Instead, the concept of preference I will use in what follows is one of preference as *considered value judgments*. On this interpretation, preferring x to y means *valuing x more highly than y* , and that this valuation is a persistent one that does not vary with extraneous factors in the way desires or do — they are values, not whims. While the psychological status of this notion of valuing is perhaps not clear — psychologists speak more of “attitudes” or “affect” towards things — it is familiar enough from folk psychology and is close to the everyday non-technical meaning of preferring one thing over another. We may ask people what they value, and we may make that inquiry in a situation most suited to eliciting considered judgments rather than spur-of-the-moment decisions (that is, the absence of stress, distorting incentives, and similar environmental obstacles to reasoning). If in such a situation we ask someone what they value, or what they most highly value — what they think would be a (most) good, right, or valuable situation, state of affairs or course of action, perhaps among certain alternatives — they may find the question hard to answer, but not *unintelligible*.

Nor would people typically think that we are asking about what they most strongly *desire* or what they would get most pleasure out of — most people realise (and bemoan) the fact they do not always desire most strongly that which they value most highly. This difference between desire and value judgments can be an instance of weakness of the will (“I know it would be better for me

⁵The well-being-based and the choice-based interpretations of the word “preference” have correlates in common uses of the word “utility.” Kahneman and Thaler (1991), recognising the difference, employ the more specific labels “experience-utility” and “decision-utility,” respectively.

if I didn't smoke, but I really want that cigarette right now") or reflect the fact that other things than what we value can drive our desires ("Of course it would be better if we gave more money to charity, but wouldn't it be nice to spend Christmas in the Caribbean this year?"). The difference between value judgments and pleasure is even more obvious — it may simply not be *pleasure* we get out of valued states of affairs, and we may even have to suffer considerable displeasure to obtain what we value most highly (think, for example, of the amount of pleasure some people regularly give up because they value politeness more highly and thus keep inviting people they hope will not turn up).

The notion of considered valuation should be unproblematic to economists, since it is formally no different from the standard binary relation used to analyse (other interpretations of) preference.⁶ That is not to say that it is necessarily complete or transitive, but nor is that an indispensable requirement for other notions of preference. The main difference from traditional economic notions of preference is that it is synonymous neither with welfare nor with choice.⁷ The concept of considered value judgment is also a central part of the theoretical equipment of moral and political philosophers.

Preference as considered value judgment is intermediate relative to the other two concepts of preference in the following sense. It is related to well-being because the property of arousing desire or producing well-being may itself lead one to value something. Yet it is different from well-being because there may be other, countervailing sources of value. Moreover, the relationship goes both ways. Sometimes the fact that an object is valued for other reasons endows it with desirability and pleasure-giving properties; getting or bringing about what you value can contribute to your well-being. Preference as considered value judgment is also related to choice because valuing something constitutes a motive for acting in such a way as to secure the valued object or bring about the valued state of affairs. Yet they also are different, since we often fail to act on the springs

⁶Broome (1999) argues that the formal apparatus of preference theory can be fruitfully applied to analysing ethical goodness and moral betterness relations.

⁷Sen (1987) employs a conceptual distinction between welfare, goals, and choice. He observes that the conventional use of the term "preference" in economic theory makes the threefold assumption that an agent's welfare is self-centred, that her goal is limited to maximising her own welfare, and that she makes choices in accordance with her goal. The terminology I propose here roughly identifies "preference" or "value" with what Sen calls "goals."

of considered value judgments due to error, psychological compulsion, weakness in the face of temptation, or moral turpitude. Here as well there may be a two-way relationship. Observing my own behaviour may make me formulate previously “latent” values (“I never realised how little I cared about the subject until I noticed how I was going out of my way to work on other topics.”)

Saying that people have preferences over comprehensive outcomes, then, I interpret as saying that people have considered value judgments that cannot be adequately represented as preferences over culmination outcomes only. I submit that this notion of preference is the one that is most relevant for social evaluation, which is a chief motivation for the present investigation. No plausible normative social theory is completely insensitive to people’s preferences. Even if a theory only minimally takes into account preferences, surely it is preference as value judgment it should take into account. In cases where the different notions of preference conflict — one policy may be more conducive to well-being, another policy more successful in bringing about what people value — a plausible social theory should choose the latter, other things being equal. The same point can be made in favour of giving people what they value rather than what they would choose.⁸

Defining preference as considered value judgment allows us to restate more succinctly the question whether preferences over comprehensive outcomes are rational. For in addition to enquiring about what a person’s value judgments are, we may also ask her to justify her position, give reasons for her value judgments, and respond to counter-arguments. This kind of reasoned scrutiny will highlight any inconsistencies in the value system, and values, like beliefs, may properly be required to be consistent in order to be rational. For example, a value system is not rationally sustainable if it fails to value the means necessary to achieve valuable ends *for no reason*. Either the value of the ends must be downgraded, or the value of the means must be upgraded, or both, or some new consideration of value must be introduced to *rationalise* the inconsistency and thereby eliminating

⁸In some theories, most saliently versions of libertarianism, not interfering with people’s choices is more important than giving them what they value. But that is typically argued on the basis of considerations of justice (so the *ceteris paribus* condition in the main text is not satisfied), or pragmatically as the best way of giving people what they value. I do not know of a theory which states that “giving people what they want” is better done by giving people what they would choose than by giving them what they value if we know what would achieve the latter. Even welfare economics has to conflate choice, value and well-being in order to claim that one should give people what they would choose (in the absence of market power and externalities).

it. By subjecting value systems to such reasoned scrutiny, we can try to discover which values the person still stands by when forced to think hard about them and resolve the inconsistencies. In John Rawls's terminology, we can investigate whether a system of valuation is in "reflective equilibrium." Naturally, I am not requiring as a condition for the rationality of a person's preferences that she actually has gone through a process of rational deliberation and self-examination. I am concerned with whether this would be *possible*; whether a set of purported objects of value and considered value comparisons *could* stand up to reasoned scrutiny.

The question can therefore be restated as follows: Can a set of objects of value that is purportedly not reducible to a set of valued culmination outcomes, that essentially contains references to the actions and processes leading to those outcomes, survive reasoned scrutiny and be in reflective equilibrium? Can a system of preferences irreducibly defined over comprehensive outcomes be free of inconsistencies that threaten the rationality of the value system? Or must claims to non-instrumental value of actions and processes on adequate reflection run into inconsistencies and turn out to be nothing but, to borrow from Bentham, "nonsense upon stilts"? The following sections will identify the way in which such value systems may be thought to be essentially inconsistent, and will explore the accounts that may potentially rescue them from the charge of irrationality.

3 Three kinds of valuation

A comprehensive outcome includes the culmination outcome or state of affairs (we shall denote these by letters X , Y and so on) and the process by which it came about (which we shall denote by A , B and so on). While a *culmination* outcome can be fully described without mention of its history, the description of a *comprehensive* outcome includes its constituent culmination outcome and all relevant features of its history. What are these features? One important aspect of the process that produces the culmination outcome is the *act of choice*, as emphasised by Sen (1997). "Who acts to bring about an outcome?" "What were the available alternative actions?" and "Was the ac-

tion intentional or not?” are all relevant questions for the evaluation of comprehensive outcomes.⁹ In what follows I shall focus on the rational evaluation of actions, although the choice act does not exhaust the relevant aspects of processes that individuals may have occasion to value or disvalue. In particular, actions could be performed according to (or in violation of) *rules* of behaviour, and the assessment of a process could include an assessment of the rightness of the rule that the process embodies. Whether or not an action conforms to a rule, then, could be relevant for the rationality of the action, or more precisely, for the rationality of valuing the action. But it is beyond the scope of this essay to discuss criteria for the rationality of rules themselves.

We shall therefore describe comprehensive outcomes by a vector (X, A) , which should be read “culmination outcome X brought about by action A .”¹⁰ I use the words “outcome” and “action” in a broad sense. A culmination outcome may be not an end-state that occurs with certainty, but a *prospect*, that is, a probability distribution over possible states of affairs. The word “action” means the set of actions or omissions or physical processes that produced a given culmination outcome, and includes random and natural events, non-intentional actions, and impersonal processes (*e.g.* market or bureaucratic processes). I shall occasionally use the word “process” to refer to all of these types of actions and aggregates of actions.

We have said that a value system may include as its objects of preference not only culmination outcomes but also comprehensive outcomes. Sometimes the preferences over comprehensive outcomes can be reduced to preferences over culmination outcomes; that is, the value system in question can be fully characterised with reference to culmination outcomes only (although it may be simpler to characterise it as defined over comprehensive outcomes). Alternatively, the value system is such that the actions which bring about culmination outcomes *must* be specified in order

⁹The term “the act of choice” should be understood to include the cases of non-action (where a culmination outcome results as a consequence of *nobody* acting) and of joint but decentralised actions (such as market processes).

¹⁰As mentioned in the main text, if *rules* of behaviour are seen as relevant aspects of processes, then this specification is not sufficient. The description of the action, to be sure, may include a reference to the rule in accordance with which the action is performed. But to fully assess the rationality of valuing an outcome-action combination (X, A) in a certain way, it is not enough to know whether the action conforms to the rule. We should also have to know whether the rule itself is rational, and this would require the rule to be specified and separately assessed. In the arguments that follow, I sidestep the question of the rationality of rules by assuming that to the extent the value system puts a premium on conformity with rules, the valued rules are themselves rational. With that assumption we may confine our attention to actions (relevantly described) and culmination outcomes.

to determine the ranks of the comprehensive outcomes in the preference order. Our enquiry is about whether such value systems can be rational. In order to clarify that question, we must now distinguish three ways in which actions may be valued.

First and most simply, actions can have instrumental value. From the point of view of rationality, this is an unproblematic kind of valuation. There is nothing irrational about valuing the means required in order to (best) achieve a valued end; on the contrary, it is *irrational not* to do so (unless one can adduce valid reasons not to value the means to a valued end, such as the possibility that they may be detrimental to other, even more highly valued ends). The instrumental value of an action is exactly the value of its instrumental efficacy; that is, the value of the uncertain prospect over culmination outcomes that is brought about by the action. I shall refer to this as the expected value¹¹ of the prospect.

Instrumental valuation of action cannot by itself sustain a system of preferences irreducibly defined over comprehensive outcomes. If actions are only valued instrumentally, then the value of the comprehensive outcome (X, A) , which we shall write $V(X, A)$, must always be equal to the value of the culmination outcome X , that is, $V(X, A) = V(X)$, if the culmination outcome follows from the action with certainty. If the culmination outcome is uncertain, then X denotes a prospect over culmination outcomes and $V(X, A)$ must be equal to the expected value of X given the performance of action A , $EV(X|A)$. If this were not the case, then either the value of action A is not all instrumental, or we would be double-counting.

Preferences defined over *comprehensive* outcomes, however, must violate these conditions. A value system that *irreducibly* contains preferences over comprehensive outcomes, must exhibit the following property:

¹¹When using the term “expected value” I am not referring specifically to the mathematical expectation of the value of culmination outcomes, just to some uncertainty-adjusted measure of the value of the possible outcomes. In a von Neumann-Morgenstern-type theory of rational choice under uncertainty, this value is indeed the mathematical expectation or the probability-weighted value of the possible outcomes: $EV(X|A) = \sum_{i=1}^n p(X_i|A)V(X_i)$, where X_i ($i = 1, \dots, n$) are the possible realisations of prospect X in a binary case. Whether the von Neumann-Morgenstern axiomatisation is the correct theory of rational choice under uncertainty is immaterial for the present argument, and nothing I say will rely on it. When I use the notation $EV(X|A)$ to denote the value of the uncertain culmination outcome (or rather prospect) X , given action A , it should be interpreted as the value of the certainty equivalent as defined by either a von Neumann Morgenstern-type valuation or some other acceptable theory of how to discount uncertain outcomes.

Process-dependence:

Holding their constituent culmination outcome constant, different comprehensive outcomes may be valued differently, depending on the process that brought it about. That is, we may have (X, A) strictly preferred to (X, B) .

In such a value system the actions themselves contribute to the value of the comprehensive outcome. For example, someone may prefer having a 60% chance of winning a fair election for an office he desires (outcome X through process A) to having a 60% chance of winning an unfair election for the same office (outcome X through process B).¹²

Moreover, the way the action influences the value of the comprehensive outcome may be different for different culmination outcomes. That is, process-dependent preferences may exhibit *non-separability*:

Non-separable process-dependence:

Holding their respective constituent culmination outcomes constant, the rank-order of two comprehensive outcomes brought about by the same process may change depending what that process is. That is, we may have (X, A) preferred to (Y, A) but (Y, B) preferred to (X, B) .

Consider what happens in the previous example if process-dependent preferences are non-separable. Suppose the expected values of the culmination outcomes are unequal, e.g. the probability of winning an unfair election is higher than the probability of winning a fair election, so that $EV(X|A) < EV(X|B)$. Notwithstanding this, the comprehensive outcomes may be ranked in the *opposite* order. In such a case, the candidate *ex ante* prefers the lower chance of winning fairly to the higher chance of winning unfairly, so $V(X, A) > V(X, B)$.¹³ In contrast, the *instrumental*

¹²Suppose both parties are equally good at cheating, so that if both cheat it does not affect the probability of winning.

¹³There is nothing unusual about this kind of reversal. A person may desire the prize allocated by a coin flip, yet prefer a fair coin to a coin that is biased in his favour. Or a person may want a certain job, yet prefer a corrupt recruitment procedure that lowers his chances of getting a job because he would rather lose a job through a corrupt recruitment procedure than because of an impartial judgment that he was not qualified for it. These types of preferences are common enough; it is their rationality that we need to investigate.

value of an action — its value *qua* instrument — is just the expected value of the culmination outcome. So if preferences over comprehensive outcomes are not reducible to preferences over culmination outcomes, then actions must sometimes be valued non-instrumentally.

The standard case of non-instrumental valuation is *intrinsic* valuation. If the process or action has value *in vacuo*, that is, apart from its consequences, then the value of a comprehensive outcome may be different from that of the constituent culmination outcome. Intrinsic valuation of process or actions is no less rational than intrinsic valuation of a culmination outcome. An action, while being a means to a valued end, may itself be a valued end. Actions could aim at *no* end and be intrinsically valued. A ballet lover, for example, will value the actions involved in performing a ballet for their own sake and not as means to anything else.

The intrinsic value of an action or process A , which we may denote by $V(A)$, is separate from its instrumental value which, as we said, is just the (expected) value of the culmination outcome, $EV(X|A)$. It is natural to think of the value of a *comprehensive* outcome as a composite of the intrinsic value of the action or process and the (also intrinsic) value of the culmination outcome. If the overall value is separable in these two component values, then we may write the value of the comprehensive outcome as $V(X, A) = V(A) + EV(X|A)$. Clearly, if $V(A) \neq 0$, we can explain process-dependence. Explaining the preference “reversals” caused by non-separable process-dependence requires that the intrinsic and instrumental values of the action combine in an appropriately non-separable way.

One example where intrinsic valuation of actions probably underlies preferences over comprehensive outcomes is *honesty*. The same (culmination) outcomes can be brought about through more or less honest actions, and — in spite of what microeconomic models often assume — people sometimes prefer to achieve their results through honest processes rather than dishonest ones. Part of this preference could be due to the instrumental effects of dishonesty (I expect to get a worse car from a dishonest car dealer than from an honest one); but some of it is surely due to a simple dislike of deception (even if I do not think I could have got a better deal from a more honest car dealer, I still dislike having been lied to when I realise the car was not as reliable as I was told).

The negative value of dishonesty may be *intrinsic*; someone may simply prefer it when people tell the truth than when they lie, even when truth-telling and lying are considered in isolation and not as means to any culmination outcome.¹⁴

Yet it would be misrepresenting preferences over comprehensive outcomes to say that a person who holds them *always* puts intrinsic value on the actions involved. When the value of a comprehensive outcome partially depends on the actions that led to it, those actions are often valued or disvalued not intrinsically, but *qua paths to outcomes*. A decision process or allocation process may be valued as a process *for* arriving at decisions or allocations. The instrumental function is then crucial; these actions or processes are *not* evaluated in isolation. On the contrary, these actions have *zero* value when we consider them in isolation from any state of affairs to which they may be the path. When we picture such actions *in vacuo*, they become empty gestures; they are mere motions that lose all their value when they have no relation with the relevant outcomes. If we have $V(X, A) \neq EV(X|A)$ and yet $V(A) = 0$ when A is considered *in vacuo*, then it cannot be *intrinsic* valuation of the action that sustains the process preference.

For example, we may value an opportunity for “voice” (i.e. an opportunity to express our concerns) in a decision process that affects us. Yet the valuation of voice is not *intrinsic*. It is implausible to think that we would put any value on voice if it were clear that it had *no* effect on the outcome whatsoever (not just because of bad luck, but because it had no causal relationship with the possible outcomes at all).¹⁵ The same point can be made about voting. Many people have a strong preference for actively participating in a vote, and not just for the culmination outcome of a certain candidate being elected. Yet they would put zero (or perhaps even negative) value on the action of putting a ballot paper in the ballot box if the election was rigged and their vote was not going to be properly counted.¹⁶ If the value of an action depends on something else and disappears

¹⁴The experimental study by Brandts and Charness (1999) finds that subjects are willing to reduce their payoffs in order to punish other subjects who have lied to them, even when the lying did not affect their own payoffs in any way.

¹⁵This is confirmed by research in social psychology. The subjective satisfaction people report to derive from voice is higher than what is attributable to the causal effect of voice in bringing about a more satisfactory decision. However, if subjects believe that there is no effect *at all* — if inclusion of voice in the decision process is a “sham” — then the reported satisfaction derived from voice disappears (and may even turn negative). See Lind and Tyler (1988).

¹⁶Note how calls for boycotting elections are based on the idea that actively participating in an election known to be rigged gives it undeserved legitimacy. Clearly the value of participation is not the same in this case as in the case

when the action is considered in isolation, then that value is not intrinsic but derivative.

In addition to actions that become worthless when considered *in vacuo*, there are some actions or processes that cannot even be defined in isolation from the ends towards which they are a means. When the action can *only* be defined as a path to an outcome, the notion of an “own” intrinsic value $V(A)$ is unintelligible. A poignant example of this possibility is the nature of pilgrimage. The proximate aim of a pilgrimage is to arrive at a holy site and carry out rites of religious devotion (the ultimate aim presumably being to fulfil one’s duty or be brought closer to God). The length and arduousness of the journey towards that goal, rather than increasing the cost, frequently *enhance* the value of the outcome.¹⁷ The experience of the journey may even outshine the culmination outcome (the arrival and worship at the holy site) in the person’s value system, and it may be of great value even if the goal is ultimately never reached. Yet a pilgrimage *must have a destination*. Random itineration with no determined goal could never be a pilgrimage; the very notion of an aimless pilgrimage is hardly intelligible. In particular, without the destination a journey could not be infused with the same kind of value that pilgrims experience. If the value of an action depends on something else and not just disappears, but is undefined when the action is considered in isolation, then again that value is not intrinsic, but derivative.

While some preferences over comprehensive outcomes are surely based on intrinsic valuation of actions, many and perhaps most are not. These process-dependent preferences value the actions in question only in conjunction with outcomes, since in isolation the actions are worthless or even undefined. Yet the value that they do put on the actions is not equal to the actions’ instrumental values. These preferences value actions *derivatively and non-instrumentally*. It is mostly this preference pattern I address in what follows, since, as I have argued, instrumental and intrinsic valuation of actions pose no special consistency problems. When they reflect intrinsic valuation of actions, such preferences are entirely legitimate from the point of view of rationality. But can a value system that endows actions and processes with *non-instrumental derivative value* be ra-

of a free and fair election.

¹⁷This example is discussed by Hirschman (1982, p. 88) who also mentions that modern sports fans practice their fandom in ways quite similar to medieval pilgrims — the further away their home team is playing, the greater the value in travelling to support it.

tional? There is something puzzling about this combination, and the next section develops two arguments why such value systems are indeed irrational. The first is based on the similarity between process-dependence and caring about sunk costs. The second argues that non-instrumental derivative valuation of actions involves an inconsistency that makes process-dependent preferences based on such valuation self-defeating in a way that is not rationally defensible. While the first argument fails, the second necessitates an account of how non-instrumental derivative valuation can be rational. The remainder of the paper discusses two such accounts.

4 Why preferences over comprehensive outcomes may be irrational

4.1 Are preferences over comprehensive outcomes like preferences that are sensitive to sunk costs?

Economics teaches that it is irrational to care about sunk costs. Sometimes it is even thought of as a paradigmatic case of irrationality. More precisely stated, the idea is that it is irrational to let the degree to which I have incurred irrecoverable costs in the pursuit of one project influence whether or not I should continue pursuing that project or rather pursue a different one. If my valuation of the actions open to me is not reducible to the expected net profits that I reap from the consequences of the actions, if in particular I disvalue the action of giving up a project I have already spent a lot of resources on even when that is the most profitable course to take, then my value system is irrational. The thought that it is irrational to let one's choices be determined by sunk costs is cemented in folk wisdom, which exhorts people to "let bygones be bygones," or not to "cry over spilt milk."

Non-separable process-dependence in preferences means that the value ranking of two outcomes can change depending on the history of how the outcomes were brought about. This could be caused not just by derivative non-instrumental valuation of actions, but also by non-separably

intrinsic valuation of actions. Non-separable process-dependence seems similar to the allegedly irrational sensitivity to sunk costs. Does this mean preferences over comprehensive outcomes are irrational as well (when such outcomes reflect non-separably intrinsic valuation or derivative non-instrumental valuation of actions)? If preferences that are sensitive to sunk costs are irrational, do preferences over comprehensive outcomes share those features of sunk-cost-sensitive preferences that make the latter irrational? An answer to this question requires us to disentangle precisely what is irrational in caring about sunk costs.

Let us first acknowledge that preferences that are sensitive to sunk costs in one way or another are ubiquitous. The fact that folklore contains admonitions against letting one's choices be affected by sunk costs bears witness to our propensity to do precisely that. We are all familiar with the difficulty of abandoning a practice or an enterprise into which we have put a lot of economic, intellectual or emotional investment, even in the face of strong evidence that it is no longer worthwhile. The scholarly literature is also well-stocked with observations of the phenomenon. A well-studied example in the context of asset investment and gambling is loss-aversion (see Kahneman, Knetsch and Thaler 1991) that can make people throw good money after bad because they are unwilling to realise their losses. More removed from the traditional subject matter of economics, Albert Hirschman proposes that a "rebound effect" can illuminate many of our social choices. "[T]ake a group of people who have experienced a great deal of disappointment in their search for happiness through private consumption: they are infinitely more 'ripe' for collective action than a group that is just setting out on that search. It is often possible to explain the choice of an amorous partner that is puzzling to outsiders on the grounds that one (or both) of the persons concerned was 'on the rebound' from another involvement that ended unhappily" (Hirschman 1982, p. 80). Hirschman's argument for taking such effects into account is not that they are rational, but rather that this is how real people behave, and it should therefore not be assumed away if we intend to predict or explain human and social behaviour: "... it may be argued that it is unsatisfactory to have the probability of the turn to public action rest on what are essentially systematic estimating biases and errors on the part of the decision makers. I do not really agree with this objection, for

... mistake-making is one of the most characteristic of human actions, so that a good portion of the social world becomes unintelligible once it is assumed away” (Hirschman 1982, p. 81).

Robert Nozick (1993) also points out the frequency with which people let their behaviour be affected by sunk costs, but highlights the possible usefulness of this tendency. “We can knowingly employ our tendency to take sunk costs seriously as a means of increasing our *future* rewards. If this tendency is irrational, it can be rationally utilised to check and overcome another irrationality.” So if I think it would be good for me to attend a number of concerts this year, and yet know that on the night of a concert, I will not be motivated to leave the comfort of my house, I can buy tickets in advance, anticipating that I will not want to have wasted the money and so will overcome the temptation to stay at home.

The ubiquity of behaviour that is affected by sunk costs, or the usefulness of such preferences, do not, of course, imply that they can be sustained in the face of rational scrutiny, and as we said, the conventional supposition in the economic theory of production is that they cannot. But why is it so obvious that sunk costs ought not rationally to matter? The claim in the case of the capitalist firm seems to draw its force from a more general formal requirement of rational preferences: Rational valuation of actions should be *forward-looking*. At the point in time when I assess the relative value of different courses of action (my own or others’), the past is fixed. Any achievable states of affairs cannot differ in their histories up to that point. All that should matter for the rational judgment of their relative values, then, should be their different futures, not their identical pasts (relative to the time of the value judgment). A rational system of value judgments, this argument would conclude, should therefore be completely describable as attributing value to current and future states of affairs. Supposing that this is indeed the general form of the specific claim that it is irrational to care about sunk costs, does it show that it is irrational to let the value of states of affairs depend on past actions?¹⁸ If we want to rescue the rationality of such preferences, we must either deny the formal requirement that rational value judgments be forward-looking in the

¹⁸More precisely: Does it show that *if* we think caring about sunk costs is irrational, then we must *also* think it irrational to have genuine process-dependent preferences? Since I have no particular stake in denying that it is irrational to care about sunk costs, I am taking that claim for granted. If it could be shown that there is in fact nothing irrational about it, then this case against process-dependent preferences would be further weakened.

sense outlined here, or we have to show that that requirement does not rule out process-dependent preferences.

Is it a formal requirement of rational valuation that it must be forward-looking? Someone might propose the following counter-argument, purporting to show that there is in fact nothing irrational in crying over spilt milk: “It is not irrational to wish that the past had been different. For example, I may prefer that the Athenians had let Socrates die of old age. I regret that Socrates was executed. What is irrational about that?” But this counterexample is misplaced because it confuses *wishing* or *regretting* with *preferring*, in the sense I have defined the latter term. Of course I may *wish* that Socrates had not been sentenced to death, without being irrational. But what does it mean to say that I wish he had not been sentenced to death? Presumably that *seen from a perspective* where the life and death of Socrates were achievable objects of value (before he was killed) I would have preferred (valued more highly in my system of considered value judgments) that he were not killed. In particular, I might say that from that perspective, I would value more highly a world in which Socrates was still alive, than one in which he was dead. Or I would value more highly the consequences of letting Socrates live than those of putting him to death. Or I would strongly disvalue the state of affairs that consists of the Athenians killing a noble man. But all of these are forward-looking reasons, as seen from the perspective in which Socrates’ life and death were still achievable objects of value. In other words, wishing that Socrates had not been killed is not a counterexample to the requirement of forward-lookingness. That requirement need only say that a rational value system should be expressible *from such a perspective* in terms of (current or) future objects of value only (in particular Socrates’ potential destinies) without reference to past valuations. The example illustrates an interesting difficulty involved in conceptualising preferences over different (counterfactual) pasts. It seems that the only sense we can give to a preference defined over past states of affairs is as a preference over current and future states of affairs, seen from the perspective in which these objects of value were conceivable future histories. Thus far from being a counter-example, the example actually reinforces the intuition that rational valuation

is forward-looking.¹⁹ Correctly understood, the claim that sunk costs should not matter, and its generalisation as a requirement that rational valuation be forward-looking, still stands.

Derek Parfit (1984, p. 165-6) has proposed the following example intended to show that we have preferences over the past. Consider a patient who is in hospital in order to undergo a very painful operation. Anaesthetics cannot be used for this type of operation, but at the end of it the patients are given an amnesia-inducing drug, so that they do not remember the pain when they wake up. When this particular patient wakes up in the hospital one morning, he cannot remember if he has actually had the operation. He asks the nurse, who knows that he is one of two patients: one who underwent a ten-hour long, very painful operation yesterday, or another who will have to undergo a one-hour operation later today. As the nurse goes to find out which of the two it is, the patient strongly prefers that he is the former. Thus, one might prefer ten hours of pain yesterday to one hour of pain in the immediate future, and Parfit proposes that most people would not think this preference irrational. But what the patient prefers is to be the patient who is *not going to undergo pain* in the future, not to be the patient who *has undergone pain* in the past, and the example confounds these two. Consider an example in which the second patient is one for whom the doctors have determined that no operation is necessary, so he will be discharged from the hospital just like the first patient. Except for their different histories, the two patients are in *identical* situations. In this case, there seems to be *no reason* to prefer being the second patient rather than the first. At the point in time where the preference is formed, it simply does not matter that one of the patients had a painful operation in the past. Any plausible reason that could be given for preferring to be one rather than the other would have to appeal to the likely future consequences of the past pain. The force of Parfit's example, then, lies not in a preference for past over future suffering, but in a preference for less future suffering rather than more. One may still, of course, *wish* or *desire* that the past suffering had not taken place, but there can be no reason to value the two situations differently *at this point in time*, since neither patient will remember any pain.

These examples do not, then, militate against the idea that rational valuation should be forward-

¹⁹Of course the history may give us relevant *information* about what these destinies are likely to be.

looking. We have argued that rational valuation of action only looks to future values, and we have accepted that sensitivity to sunk costs is irrational. So if the irrationality of caring about sunk costs is not to entail the irrationality of process-dependent preferences in general, it must be a mistake to think that forward-lookingness incriminates those preferences in the way it does sensitivity to sunk costs.

I propose that the argument that process-dependent preferences are irrational because they are like caring about sunk costs rests on a confusion. Once the confusion is clarified the argument cannot be sustained. The clarification consists in distinguishing the requirement that reasoned valuations of actions should be forward-looking — they should only take into account achievable (presently or in the future) objects of value — and the claim that *future values* cannot themselves be partly determined by that path taken to the valued objects. The requirement of forward-lookingness applies to the relationship between actions and what valuable things they achieve. It does not imply anything about the temporal relation to the *source* of those values. What makes sunk costs a special case is that it assumes that the metric of value is profits (or more generally, some *net* sum of benefits and costs, which need not necessarily be monetary). Once an outcome is described in terms of the profits (or net benefits) it yields, there is nothing else left to enter the calculation of value, and in particular, the *gross* costs incurred in the process do not affect the value of the outcome unless it affects future profits. We can illustrate the distinction in terms of the simple formalisation introduced above. If the only thing that is valued is profits, then the value of a comprehensive outcome can be written $V(X, A) = E(\pi(X, A)) = E(\pi(X)|A)$ where π denotes net profits. That is, the value of the comprehensive outcome (X, A) is the expected value of the profits generated in prospect X brought about through action A . This expected value — since profits are part of the culmination outcome — is just the expected profits in X , where the expectation takes into account the conditional (causal) probability of the possible realisation of X , given A . So the total value of the comprehensive outcome is just the expected value of the culmination outcome, $EV(X|A)$.

If sensitivity to sunk costs is irrational, then, it is not because the value of a state of affairs rationally ought not to depend on the past, but because of the choice of a specific metric of value

that is defined independently of the gross costs incurred along the way. It could be perfectly rational for an agent to be sensitive to sunk costs if she values other things than profits and if sunk costs are an indicator of other (positive or negative) values. In such a case, choosing to stick with a project into which a lot of resources have already been put may achieve *other* values even if it involves “throwing good money after bad.” The assumption that only profits matter may make perfect sense for the firm in neoclassical producer theory, but it is not a generalisable feature of systems of rational value judgments. In the general case, the value of a comprehensive outcome may depend on the action (or omission of action) that brings it about as well as the expected value of the culmination outcome.

Why might the value of a state of affairs depend on its history? Consistency over time may itself be an important part of a person’s value system. Such consistency is most obviously important in those value systems which put a premium on integrity. Integrity requires acting in a way that gives some unity to the choices a person makes over the course his life, and consistency of choices over time may give a valued sense of completeness to a person’s identity. It is a criterion against which choices of various kinds are often measured; they include moral choices in particular, but also choices that involve social and intellectual commitments even without being deeply moral.^{20,21} Thus a history-dependent metric of value may be necessary to understand value judgments reflecting a persistent commitment of a personal, political or intellectual kind. These commitments are an important part of some of the most important choices we can make; they are displayed when someone remains with their old and sick spouse out of loyalty even when all remains of earlier love have withered; when a captain chooses to go down with his sinking ship; or when an officer who is a prisoner of war has the opportunity to be released but refuses to do so

²⁰See Williams (1973) for an insightful discussion of personal integrity. We have mentioned how Robert Nozick points out the usefulness of such path-dependent preferences. Although he does not elaborate on their rationality, he also stresses that they are important components of our personal identity: “We do not treat our past commitments to others as of no account except insofar as they affect our future returns. . . and we do not treat the past efforts we have devoted to ongoing projects of work or of life as of no account (except insofar as this makes their continuance more likely to bring benefits than other freshly started projects would). Such projects help to define our sense of ourselves and of our lives” (Nozick 1993, p. 22, original italics).

²¹Note the value that is often placed on intellectual integrity, which demands that one should (to a certain degree) attempt to make one’s beliefs about something be consistent with the beliefs that one already holds or has previously stated about other things.

as long as his subordinates are not released with him. On a more prosaic level and to return to the central case of this section: integrity may be a good reason to be sensitive to sunk costs. If I have put a lot of resources and effort into my project, my value system may charge me with a special responsibility to “see the project through,” even if it is no longer worthwhile in the eyes of someone who has not invested the same effort.

In none of this have I claimed that non-instrumental valuation of processes is rational or irrational. What this section has argued is that *if* it is irrational, then that cannot be because of a purely formal requirement of forward-lookingness. Forward-lookingness, I argued, may mistakenly be thought to be why sunk-cost sensitivity is irrational. If forward-lookingness is indeed a formal requirement of rational valuation of actions, however, then it must be understood to require that only current or future *values* matter, not that those values themselves must be determined without reference to the past. Such a requirement would rule out integrity or consistency on purely formal grounds, and it would be a strange theory indeed that ruled consistency in violation of formal rationality! Any argument that it is irrational to have preferences over comprehensive outcomes must therefore be based on identifying substantive inconsistencies. It is to that task we now turn.

4.2 Are preferences over comprehensive outcomes like “future Tuesday indifference”?

We have seen that the sunk cost analogy fails, and that the purely formal feature of non-separable process-dependence cannot by itself incriminate preferences over comprehensive outcomes at the tribunal of rationality. Intrinsic valuation of action, even if non-separable, escapes the irrationality charge. But we said that there are two kinds of action valuation that could cause process-dependence. In addition to having intrinsic value, actions or processes may have value that is *derivative* because conditional on their instrumental function as actions or processes *towards* culmination outcomes. Yet this derivative value (in the case of preferences irreducibly defined over comprehensive outcomes) is not merely instrumental, since it cannot be accounted for solely in terms of the instrumental efficacy of the action. This is a problem. Such a system of preferences

is liable to the charge that, *for no reason*, it may prefer something less highly valued above something more highly valued. If that claim can be sustained, such a preference pattern is irrational: It exhibits an inconsistency that cannot subsist in reflective equilibrium. While the failure of the sunk cost argument leaves unscathed process-dependent preferences reflecting intrinsic valuation of actions, we have yet another charge to answer with respect to preferences over comprehensive outcomes that are based on derivative non-instrumental valuation of actions.

It is useful to consider an analogy of how inconsistent values run afoul of rationality. Derek Parfit's (1984, p. 123-4) case of "future-Tuesday-indifference" illustrates what the consistency requirements on rational preferences must minimally be. Parfit asks us to imagine a person who puts high negative value on physical pain, yet is indifferent today to pain that will be inflicted on him on a future Tuesday. He is adamantly non-indifferent to pain inflicted on other future days, and during a given Tuesday, he dislikes pain just as much as he does on other days, and he knows that on a given future Tuesday he will strongly dislike the pain that he could now make choices to avoid. Other things being equal, he will therefore today prefer a very painful treatment on a future Tuesday to a less painful treatment on a future Wednesday even though he generally prefers less pain to more. He has no religious or other reason to justify his preference; it is simply a fact about him. Is this preference rational? We must agree with Parfit that "preferring the *worse* of two pains, *for no reason*, is irrational" (Parfit 1984, p. 124, original italics). Future-Tuesday-indifference is irrational because it is a specific instance of a more general inconsistency: It puts something avowedly less valuable (in this case, more pain) over something avowedly more valuable (in this case, less pain), *for no reason*. That inconsistency cannot survive rational scrutiny. A value system in reflective equilibrium does not exhibit future-Tuesday indifference.

Preferences over comprehensive outcomes (that are not based on intrinsic action values) seem to suffer from the same inconsistency. If the derivative value of actions and processes is more than (or less than) the value of their instrumental efficacy, then situations are possible in which the values of comprehensive outcomes are ranked differently from the values of the constituent culmination outcomes. Culmination outcome *X* may be valued more highly than culmination outcome *Y* *in*

vacuo, and yet comprehensive outcome (X, A) may be valued less highly than comprehensive outcome (Y, B) . This value system sometimes puts Y above X (when they are achievable through actions A and B , respectively) even though on its own terms X is more valuable than Y . Can this be rationally sustained? The reason for the change is the value of the actions A and B . But suppose that those actions are not *intrinsically* valued or disvalued, but rather *derivatively* valued or disvalued. And their derivative value would then be contingent on their instrumental function *qua* ways of bringing about X or Y , without being reducible to their instrumental value. If their value ultimately derives from the culmination outcomes at which they instrumentally aim, then how could that derivation make the less valued culmination outcome come out “on top” in the ranking of comprehensive outcomes? In other words, how could the derivative value of a *means* to a culmination outcome, an action that is valued *qua* means, be different from its instrumental value, *i.e.* the value of its instrumental efficacy? Such preferences seem to rank the less highly valued above the more highly valued *for no reason*. If that is right, then they must indeed be irrational. A value system which promotes actions that undermine the source of their own value (that lead to less valued culmination outcomes) is what Parfit calls directly self-defeating.

Consider the value system of an *aesthetic mathematician*.²² When she evaluates a proof, her main criterion is whether it is complete and correct. Yet she may also care about the beauty and elegance of the proof. Suppose this mathematician is not an aesthete in other areas of her life; beauty and elegance is not something she values by itself. In particular, she would place zero value on an elegant string of equations that served no purpose. The beauty she values is a *purposive* beauty: It is a property of the *way in which the outcome is achieved*, of how a mathematical statement is proved. The beauty of the method of proof, therefore, is so intimately linked with the result that the proof’s aesthetic value is derivative. It is also not merely instrumental, because two different proofs of an identical result could have different aesthetic value. Does this mathematician have a rational value system? Suppose she is faced with two proofs of the same result. One is long, convoluted, but complete. The other is short, elegant, and dazzling, but is not complete. I submit

²²I owe this example to Nien-Hê Hsieh.

that to be consistent with the fact that the aesthetic value is derivative, she would have to value the more complete proof more highly. Valuing the incomplete proof above the complete one is irrational, unless she can bring in other factors to justify such a preference. Letting the *derivative* aesthetic value of the inferior proof outweigh the higher intrinsic value of the complete proof, *for no reason*, is a self-defeating pattern of preference, because it values an action that undermines the source of its own value.

This immediately points to a special case of derivative non-instrumental valuation of actions that is not self-defeating in this sense. If the value system has a *lexical* structure, then it could value actions derivatively and non-instrumentally without being inconsistent.²³ In such a system, the value of a comprehensive outcome would first depend uniquely on the value of the culmination outcome (in this case, the correctness of the result). Only if two culmination outcomes were equally valuable *in vacuo* would the action or process (in this case, the elegance of the method of proof) be brought into the value calculation to break the tie. If the mathematician gave lexical priority to the preference for more correct and complete theorems over the (derivative non-instrumental) preference for more elegant proofs, her value system would not be self-defeating.

In general, however, there is no reason to expect that people who value comprehensive outcomes differently from culmination outcomes (when the preference is not due to intrinsic valuation of the action) do so in a lexical fashion. The fact that the problem is avoided in the special case of lexicographic preferences simply serves to underline the seriousness of the irrationality charge. A value system reflecting derivative non-instrumental valuation of actions can be in reflective equilibrium if only it is retuned so as to essentially *ignore* that kind of valuation, except when those values are needed as tie-breakers. This condition, of course, all but rules out process-dependence and completely excludes non-separable process-dependence. The charge that non-instrumental derivative valuation of actions is self-defeating must therefore be rebutted if we are to show that such preferences can be rational outside of the very special case of lexicographic preferences.

The irrationality of letting mere means to valuable ends take on value that is not warranted by

²³Richard Tuck suggested this possibility.

their success in bringing about those ends has been a common theme with moral thinkers. The prime example is valuing wealth for more than its ability to secure other valued things. Adam Smith cautioned that “[m]any a poor man places his glory in being thought rich, without considering that the duties (if one may call such follies by so very venerable a name) which that reputation imposes upon him, must soon reduce him to beggary, and render his situation still more unlike that of those whom he admires and imitates, than it had been originally” (*Theory of Moral Sentiments*, I.I.3). Just like letting wealth borrow value from the desired objects it is thought to secure may be counterproductive to the actual fulfilment of one’s desires, so may letting actions and processes derive value non-instrumentally from culmination outcomes be counterproductive to the achievement of those outcomes. To bring such a value system to reflective equilibrium something would have to give. Either the valuation of actions or of outcomes would have to be adjusted.

A frequent preference pattern that falls under this accusation of being self-defeating is the value of participating in the act of voting. People vote in order to elect, say, the president. For many, voting is an activity of deep meaning and importance. Yet the value of the act itself is derivative; it is not the mere motions of putting the ballot paper in the ballot box that count. The importance of voting derives from the fact that it is a process by which the president is selected; the act of voting has no value in a rigged election, as we have already mentioned. Thus the knowledge, or at least the belief, that one’s vote is counted is crucial for the value of voting. In this sense the value of participation in an election is derivative; it is conditional on the instrumental function of the vote. But there exist alternatives to voting that are equally efficacious in carrying out that instrumental function. In particular, there is the possibility of vote-trading. The instrumental efficacy of my participation is identical to that of convincing somebody else to vote for my candidate who would not otherwise vote, and almost identical to that of convincing somebody else not to vote who would otherwise vote for the opposing candidate (depending on the number and support of the candidates and the format of the election). And if a vote for my candidate is more likely to be pivotal in a different constituency than my own, then the instrumental efficacy of engaging in vote-trading is *greater* than of voting for my own candidate: I can increase the probability of my

preferred electoral outcome if I can trade my vote against someone else's more pivotal vote for my candidate in another constituency (if I trust that my counterparty will carry out his or end of the deal). This was the reasoning behind the internet-based vote-trading scheme whereby Greens in Florida would vote for Al Gore in return for Democrats in less pivotal states voting for Ralph Nader in the 2000 U.S. presidential election. Those who promote vote-trading schemes can argue that since the value of voting is contingent on its instrumentality in selecting a candidate, there is no reason to value the act of voting differently from its (expected) instrumental achievement. If we can all save time and effort by trading our commitments not to vote, or if we can all improve our expected culmination outcomes by trading votes across constitutencies, then assigning non-instrumental value to voting in a way that prevents a Pareto-improvement *for no reason* is self-defeating, hence irrational. A citizen who prefers personally casting his vote for his preferred candidate to engaging in instrumentally equivalent or superior vote-trading, one may argue, is like the mathematician who sacrifices the completeness and correctness of the proof for the sake of its elegance. Yet many would not just reject this argument for vote-trading, but find it offensive. They would respond that voting is a duty, or that considering the act of voting as something that can be traded profoundly misses the value of voting. They would not be indifferent between voting and (instrumentally equivalent) vote-trading. Indeed vote-trading is illegal in most democracies (and the aforementioned scheme was banned).

To show that such preferences can be rational, it is necessary to explain how the derivative value of actions — which is contingent on their instrumental function — can rationally be different from their mere instrumental value. I shall offer two accounts, which both rescue the rationality of such preferences, but in fundamentally different ways, and with divergent consequences for the role of individual preferences in social evaluation. The first account retains the focus on instrumental rationality, but distinguishes between direct and indirect instrumentality. The value of actions may not be reducible to their instrumental efficacy, but the assigning of such seemingly irrational value to actions may itself be instrumentally most valuable because it best achieves valuable culmination outcomes. Preferences over comprehensive outcomes are, on this account, an instance of *rationality*

irrational preferences. In the next section I shall discuss what would have to be true for this account to be correct, and will argue that the nature of those conditions makes it implausible that preferences over comprehensive outcomes should be taken into account in social evaluation exercises.

The second account, unlike the first, accepts that non-instrumental valuation of processes may be rational even if it is not justifiable in terms of indirect instrumentality. This can be because derivative values need not be merely instrumental values. Following Nozick (1993), section 6 discusses how one object can derive or impute value from another not only through instrumental or causal connections, but also through evidential or symbolic relations. Since evidential or symbolic relations may require instrumental functionality but not be reducible to it in intensity, preferences based on them can be rational if evidential or symbolic value transfers are. While not unproblematic, I shall argue that this account better represents preferences over comprehensive outcomes than does indirect instrumentality. One implication of this view will be that preferences over comprehensive outcomes must have a much more central role in social evaluation than is typically granted them in economics and other social sciences.

5 Indirect instrumentality

I showed in the previous section how preferences over comprehensive outcomes can be directly self-defeating. When the derivative value of actions is not equal to their instrumental value, the value system will sometimes promote actions that frustrate the achievement of what is ultimately most highly valued. If preferences over comprehensive outcomes have this implication, then they are irrational. There is an answer to this argument which preserves the focus on instrumentality as the only source of derived value. The answer is this: “A value system that is not sometimes directly self-defeating in the above sense, is indirectly self-defeating. Even if we only ultimately value culmination outcomes, we will more efficiently bring about those outcomes if we also attribute derivative non-instrumental value to actions and processes. Even if derivative non-instrumental preferences for actions are sometimes directly self-defeating, such preferences more efficiently

promote the things that are ultimately valued than a value system which avoids being directly self-defeating.” In other words, the *indirect* instrumental efficacy of non-instrumentally valuing actions and processes outweighs the occasional *direct* instrumental inefficacy. Therefore, the overall most rational value systems may be locally irrational. Preferences over comprehensive outcomes, on this view, are an instance of *rational irrationality*.

Rational irrationality is a well-understood concept in philosophy and in the social sciences. Two much studied examples are the benefit of being able to make threats or promises that are costly to carry out (the problem of “incredible threats”), and the benefit of committing to a future action which at the time when performance is called for one will not want to perform (the problem of temptation or “time inconsistency”). In both of these cases, it is overall more rational to do what is irrational in a local context. The move from local to global instrumental rationality is analogous to the change of focus from individual acts to rules or dispositions within utilitarianism (and consequentialist moral theories more broadly). In utilitarianism, the question is whether, for an action to be right, the action itself should maximise overall welfare — as in direct utilitarianism or act-utilitarianism — or the action should reflect the *disposition* that maximises overall welfare — as in indirect utilitarianism or rule-utilitarianism. In the study of instrumental rationality, the question is whether, for the derivative valuation of an action to be rational, the action itself should lead to the most valued culmination outcome, or the value of the action conform with the overall value *system* the believing in which leads to the most valued culmination outcomes. We may follow the utilitarian labelling and distinguish between *direct* instrumental rationality of actions or simply act-rationality, and *indirect* instrumental rationality of actions (instrumental rationality of value *systems*, or of the *disposition* to value certain things) or rule-rationality.²⁴

Just like in utilitarianism there is a debate as to how there could be a difference between act- and rule-utilitarianism (what Davis Lyons (1965) calls “extensional equivalence”), so the question arises here how there could be a difference between act- and rule-rationality. We have already mentioned one way in which what is locally irrational can be globally rational. Local irrationality

²⁴Robert Aumann used the terminology of act-rationality versus rule-rationality in his PELS lecture at Harvard University in 2002.

may serve as a commitment device. This makes it rational sometimes to suspend rationality, if possible. But preferences over comprehensive outcomes do not in general seem to fall into this category. It is hard to find many examples of non-instrumental valuation of actions serving as a commitment device against short-term temptations in order to secure long-term valuable outcomes. The best may be the existence of *reciprocity motives*. Some experimental studies suggest that people are more averse to accepting unequal distributive outcomes that result from intentional actions than those resulting from allocations by random devices (Blount 1995). If people ultimately care about getting the highest payoff for themselves, the disposition to punish ungenerous people could result in locally irrational rejections of Pareto-improving options, but globally encourage more generous behaviour by others. Punishing random devices, on the other hand, would be both locally and globally irrational. The problem with this account is that it fails to explain how it could be rational for people to act like this even in one-shot anonymous games with subjects they will never meet again. One would have to make heroic assumptions about the complexity of people's instrumental calculations and concern with very improbable small-stake events (the chance that they later interact with someone who knows how they behaved in the experiment) for this argument to work.²⁵

The best case for extensional non-equivalence between direct and indirect instrumental rationality is the existence of imperceptible incremental effects of actions on outcomes. This phenomenon is an instance of the ancient sorites paradox of imperceptible differences (also called Wang's paradox). Suppose I want to make a pile of sand. How many grains of sand do I need to accumulate? The addition of a single grain of sand will not make something that is not a pile into a pile,

²⁵Another way in which irrationality may be thought to be rational is in the presence of cognitive and informational limitations. We do not always know what consequences our actions or those of others may have, because we do not have all the relevant information or because we do not have the time or the computational resources to process it. The best we can do is to rely on rules of thumb. Since rules of thumb will sometimes tell us to act differently from what a thorough calculation in the presence of a better information would conclude, we are sometimes forced to make suboptimal choices. But this is not rational *irrationality*. Subject to the informational and computational constraints, following rules of thumb is rational even locally. But in cases where it is known that the effect of an action will be to achieve the most highly valued culmination outcomes less well, following rules of thumb is both locally and globally irrational. And it is precisely in these instances that the question of how rational non-instrumental derivative valuation of actions is has relevance. In the easy cases, we do not need to pay particular attention to such values, since they do not differ from a purely instrumental perspective.

yet with enough grains of sand I obtain a pile. This causes problems for maximisation exercises. If I value both a pile of sand and I disvalue effort, then my most highly valued end-state would be the one where I have made a pile at the minimum possible exertion of effort, but there is no solution to the minimisation problem of how many grains of sand I should pile up. Direct instrumental rationality fails in such cases, and an agent's goals may be best served by adopting rules of behaviour that seem locally irrational. Whenever the effects of single actions added together are in this way different from the effects of a general practice of that action, there is extensional non-equivalence and indirect and direct instrumental rationality may diverge.

An example shows how the sorites phenomenon may make it indirectly instrumentally valuable to value actions not merely (directly) instrumentally. Consider a procedural preference for *politeness* — that is, a valuation of polite actions and processes by more than their instrumental efficacy in producing valued culmination outcomes. The instrumental value of politeness is clear: Treating others politely will often make them more likely to do what we desire. But this is not always the case. Sometimes one has to be impolite in order to reach the most highly valued (culmination) outcome out of those available. A pure instrumentalist should not care about being polite in those cases (nor, presumably, should he care about politeness in the actions of others unless it affects the achievement of the outcomes he values). But it is likely that someone who is known to be polite only when it helps him get what he values may be disliked by those who know him (certainly they will not value his politeness as much as if they did not think he was only polite for instrumental reasons), and this may in turn hinder the achievement of his most highly valued states of affairs. On the other hand, nobody will mind an occasional lapse of civility (“I really had no other way of getting rid of the telemarketer than to rudely hang up!”). The instrumentally rational thing to do is to balance the positive and the negative marginal effects of being rude slightly more frequently. But this is a sorites problem. A marginal increase in rudeness has *no* negative effect, even though a large increase does. Direct instrumentality, applied to each single act separately, leads to worse outcomes than are achievable. Indirect instrumentality, on the other hand, would tell the person to value politeness non-instrumentally (although perhaps not too strongly) and thus overcome

the sorites problem. Even though this non-instrumental component of the person's value system promotes actions that seen individually are instrumentally irrational, it leads to better outcomes overall than does a value system that attributes to each action its direct instrumental value, because it avoids producing the frowned-upon personality trait. Non-instrumental valuation of actions, even when it is *directly* self-defeating, may perform better than directly instrumental valuation of actions, which is *indirectly* self-defeating.

The presence of sorites problems in the effects of actions thus provides grounds for appealing to the indirect instrumental value of not valuing actions merely (directly) instrumentally. But if this is the only way in which preferences over comprehensive outcomes can be rational, then we should conclude that the process-dependent part of people's preferences should not be given the same importance in social evaluation as more narrowly defined preferences over culmination outcomes. There are two reasons for this. The first reason is that indirect instrumentality rescues their rationality only in a very limited sense. As discussed in the introduction, irrational preferences are not and should not be treated in the same way as rational preferences in social evaluation. While they ought perhaps not to be ignored altogether (liberalism and caution both suggest that one should avoid as much as possible to discriminate between different kinds of preferences), preferences that can be shown to be irrational do have less moral significance for collective decisions. If they are based on error, for example, a concern for people's preferences may itself advocate discounting values that are not sustainable under reasoned scrutiny. Indirect instrumentality only rescues the rationality of process-dependent preferences in the sense that it is useful to have them, and that if one has the ability to cultivate them, it is rational to do so. But this cannot be done through a rational process. That becomes clear when we note that rational preference formation has analogies to rational belief formation. It may be instrumentally useful to believe in two inconsistent propositions. The fact that it is useful for someone to have a certain belief, however, does not provide a *reason* for actually having that belief. It may *induce* that belief through a psychological or evolutionary process, but the belief cannot survive rational scrutiny in the absence of other reasons. Similarly, the fact that it is useful for someone to hold something valuable is not in itself a reason

for actually holding that value. It may induce that valuation through a nonrational psychological or evolutionary process (which it may be rational to try to trigger), but the value itself cannot survive rational scrutiny in the absence of other reasons. “Because it is useful for me to think so” is not a rationally acceptable basis for holding either a belief or a value.

The second reason why preferences rationalised by their indirect instrumental value deserve less than full merit in social evaluation has to do not with the formation of the preferences but with their justification. The presence of imperceptible effects (on the achievement of valued culmination outcomes) is what is claimed to fend off the threat of inconsistency by making locally irrational preferences globally rational. Given certain constraints (the imperceptibility of the effects of actions), *individuals* should rationally cultivate such preferences if *they* want to maximally achieve what they ultimately value most highly. But it is extremely implausible to think that the government (or whichever other collective entity is the focus of the social evaluation exercise) faces the *same* constraints on how *it* can maximally achieve what people most value. The reasons why it is indirectly useful for individuals to value actions and processes non-instrumentally do not provide the government with reasons to adopt the same non-instrumental values. There may be other imperceptible effects of the government’s actions which make it indirectly useful to value certain collective actions non-instrumentally. But this is not an argument that there should be a fit between individual and collective preferences over comprehensive outcomes in the ways normally required by theories of social evaluation in the case of standard preferences. On this account, preferences over comprehensive outcomes, at least when based on derivative non-instrumental valuation of actions, are necessarily less important than preferences over culmination outcomes, regardless of how exactly the latter are incorporated in social evaluation.

I propose that we reject this account as empirically false, or at least unrepresentative of most cases of process-dependent preferences. The view that non-instrumental valuation of actions and processes is rational because and insofar as it reflects instrumentally optimal rules or dispositions sees the features of the world that make such dispositions useful (the presence of imperceptible effects) as *constraints*. Extensional non-equivalence is an *obstacle* to the maximal achievement

of the valued culmination outcomes. If the globally best value systems are ones that sometimes prefer the locally suboptimal actions, then sometimes “getting it wrong” is the price we have to pay to “get it right” overall. But this fundamentally misrepresents most cases of preferences over comprehensive outcomes. Agents with such preferences do not see them as second-best solutions to frustrating obstacles to their goals. Instead, this feature of their value systems — values about how things are done — is likely to be a deep source of identity and may even be what gives many valuable culmination outcomes *their* value. That such agents do not see their non-instrumental values as solutions is shown by the fact that they would not appreciate a change that made those values redundant. If extensional non-equivalence ceased to exist (by an acute improvement in our ability to perceive previously imperceptible effects), non-instrumental valuation of actions would lose its indirect instrumentality. On this account, rational agents should immediately revise their preferences and cease to attribute derivative non-instrumental value to actions. This flies in the face of the phenomenology of such preferences. Agents who hold them would not appreciate an opportunity to make the procedural component of their values redundant. That component frequently provides a source of meaning to their actions, and if people no longer had occasion to exercise those value judgments, they would most likely experience it as a loss and impoverishment, rather than feeling relieved by the removal of a constraint on the maximal fulfilment of their values.

6 Non-instrumental derivative value

As we discussed above, the imputation of value from ends to means is unproblematic. It may even be, as Nozick (1993) suggests, that any theory of rationality must have room for instrumental rationality and that that may be the only common core of all plausible theories of rationality. Nozick takes the obviousness of value derivation through the causal (instrumental) relation as a springboard for pointing out that value is also imputed through other channels than the causal one. Through their causal/instrumental relation to outcomes that yield “utility,” actions have “causal expected utility.” In addition to the instrumental channel, Nozick claims value may also be derived through an evidentiary or a symbolic relation, endowing actions with two other forms of derivative

value, *viz.*, “evidential expected utility” and “symbolic utility.” What enables actions to take on these non-instrumental values is that in addition to (and partly independently from) bringing about valued states of affairs, the actions come to represent or “stand for” other valuable actions or states of affairs. This relation of representation could be an evidential relation — the performance of the action *gives evidence for* or *indicates* that the desirable state of affairs obtains — or a symbolic relation — the performance of the action *symbolises* the desirable state of affairs. The classic example of evidential utility is the Calvinist belief that earthly success was evidence of divine grace, a belief which endowed worldly riches and the actions conducive to them with more value than they had either intrinsically or instrumentally in an otherwise ascetic value system. A more contemporary example may be how consumer goods advertising associates the consumption of certain goods with certain lifestyles or types of people. This can be seen as creating “evidential utility” by making the consumption of certain goods indicate that one is a certain type of person, as when Coca-Cola encouraged consumers to drink its product with the slogan “Be sociable.”²⁶ A similar phenomenon is that of social climbing, in which people adopt activities and ways of being that identify them (to themselves or others) as being part of a more prestigious or admired social group.

Applying the idea of evidential value to the context of evaluating actions, we may say that certain actions have more derivative value than their causal efficacy warrants because they constitute *evidence* of desirable culmination outcomes. Return to the example of voting and consider the collective dilemmas involved in voting and activities like it. In an election it matters that enough people vote for their desired candidate to win, but if even more people vote, it does not make a difference to the culmination outcome. Why should I personally bother to go and vote if there is a negligible chance that my vote will be pivotal? A common retort is “What if everybody thought like that?” But my failure to vote does not *cause* other people not to vote. Nevertheless, my decision whether or not to vote might constitute partial evidence to me of what other people in my situation will decide to do. This could endow my voting with “evidential expected utility” over and above

²⁶This example is cited by Alexander Schuessler in his interesting comparison of the parallel evolution of soft-drinks marketing and presidential election campaigns (Schuessler 2000, chapter 5).

its causal expected utility (which is often considered negligible by rational choice theorists).²⁷

The second way in which an action may represent another action or a state of affairs is by *symbolising* it.²⁸ The performance of the action does not provide evidence for the valued outcome, but it comes to stand in its place. Consider the ample use of symbols in war and political contexts. Why did the American soldiers who took Baghdad drape the statue of Saddam Hussein with the flag that had been waving over the Pentagon on September 11th, 2001? Of course such actions often have purely instrumental value as well — boosting fighting morale, for example — but that is rarely their main motive. Rather, these actions are worth performing because of their powerful symbolic content. Moreover, even when the action is done for instrumental reasons, the symbolism must often work independently in order for the desired instrumental effect to follow (such as in this example). The flag in war *symbolises* the highly valued objects that the soldiers are fighting for: the security or prosperity of one's country, freedom from occupation, humanitarian goals. The draping of the statue in Baghdad symbolised a U.S. victory over terrorism that in reality is evanescent.²⁹ An action that is worthless *in vacuo* takes on derived value by symbolising something that is intrinsically valuable.

One application of this phenomenon to our case of process-dependent preferences is when the *intention* to bring about an outcome endows an action with more value than its causal efficacy warrants. As mentioned above, experimental game theory has found that subjects are much more willing to reject unequal distributive outcomes when those outcomes result from other people's actions compared to random allocation devices, even in one-shot games. Similarly, it is easy to think of cases where my influence on the likelihood of a desirable (culmination) outcome is minute, yet it matters that “I do what I can.” Why is this? One interpretation is that my action,

²⁷Nozick (1993, chapter 2) offers an insightful discussion of the role of evidential utility in Prisoner's Dilemma-type situations.

²⁸Closely related to the notion of symbolic value is that of expressive value. The distinction is not important for the present argument, and we may simply think of the expressive as a subset of the symbolic (perhaps the explicitly communicated subset of the symbolic, which also includes non-expressed symbolism).

²⁹Of course to the Iraqis it symbolised something quite different — defeat, occupation, colonisation — and was consequently perceived as a humiliation. An in itself instrumentally inconsequential action took on a hugely negative symbolic value and in virtue of that also became instrumentally important. The flag was swiftly removed as the difference in values dawned on the US soldiers.

by so to speak embodying my intention or desire to bring about the outcome, comes to symbolise the intended outcome and thus “borrows” some of its value in a way that a non-intentional process does not. This is one reason why people attach importance to *agency* and not just (culmination) outcomes. It matters not just that my candidate for president wins, but that *I* vote for him or her. Suppose I think that my candidate will be elected no matter what I do. Even though my vote is not going to be pivotal, it may matter to me that *I* am among the voters who are collectively instrumental in electing our favoured candidate. Suppose, alternatively, that I think my candidate will not be elected no matter what I do. Even when I thus have no causal effect on the outcome of the election, my vote itself may *symbolise* my favourite candidate’s victory and thereby take on derivative value that makes it worthwhile for me to vote.³⁰ This account also explains why many people have a strict preference for voting rather than dissuading supporters of other candidates from voting. The instrumental effect of the two actions is the same, but the symbolic value is very different.

In both cases — evidential and symbolic value — actions or processes “borrow” value from the valuable objects they represent through evidential or symbolic relations. The *possibility* of value transfers through these connections is no more mysterious than the imputation of value from ends to means through the instrumental relation.³¹ There is a question, however, about the *rationality* of evidential and symbolic value transfers. If such transfers are rationally sustainable, then we can also explain how actions may have derivative value that is contingent on instrumental function but not reducible to instrumental efficacy. This puzzling phenomenon is rationally possible if the instrumental function itself creates evidential or symbolic connections, which in turn allow non-instrumental derivative value to be imputed from the represented values.

Are value transfers through relations of representation defensible under reasoned scrutiny? In his own treatment of evidential and symbolic utility, Nozick uses these concepts mostly in a psychological sense. We are so constituted that we get satisfaction from objects that represent other

³⁰Schuessler (2000) outlines a formal theory of the role of symbolic values in voter behaviour.

³¹Which does not mean that it is not mysterious. Nozick (1993) points out that it is not clear how we are able to make value “travel back” in this way. My argument here merely needs to show that symbolically and evidentially derived value can be just as rational as instrumentally derived value.

desired objects. This is a psychological claim, and saying that value travels back from outcomes to actions through evidential or symbolic relations is making a statement about psychological mechanisms. For our question, however, establishing the existence of these psychological mechanisms is insufficient, since, as we argued above, psychological sensations of satisfaction are neither something to which the consistency requirements of rationality can usefully be applied, nor are they the appropriate basis for social choice evaluations. Instead, we must use the concepts of evidential and symbolic value in the different sense of *reasoned value judgments* about actions based on those actions' evidential or symbolic relation to (other) valued actions or outcomes.

The argument why non-instrumental derivative valuation of processes is irrational was presented in subsection 4.2. If actions and processes have a derivative value that is different from their instrumental value, then a person may value actions that undermine the very source of those actions' derivative value. This charge of inconsistency can be levelled against any derivative valuation that is not merely instrumental and therefore applies to both symbolic and evidential value transfers. When actions take on derivative value by *representing* a valued state of affairs over and above the value they have in virtue of *causally contributing* to its occurrence, then it is possible that the most highly valued actions available to a person, or the most highly valued processes, are such as to make the ultimately (non-derivatively) valued state of affairs *less* likely to occur than it could be.

We may continue with the example of voting to illustrate the inconsistency. Some voters cast their ballot for marginal parties that have no chance of gaining a seat; these are “wasted” votes from the point of view of their efficacy in getting one's most preferred candidate elected. Sometimes such voting is driven by other instrumental motives (such as helping the candidate secure funding or media attention in the next election) or by indifference among the non-marginal candidates, but at least in some cases, voters simply want to vote for their preferred candidate without aiming at any sophisticated oblique instrumentality.³² One reason why voters may “waste” their vote is for

³²The claim is often made that *no* merely instrumental motives can make voting rational, because of the minute probability that any one voter will make a difference to the outcome. There must be some value attached to the act of voting itself (but as I have argued, this value may be conditional on the instrumentality of voting). The argument I make in the main text does not require such a strong view, it simply focuses on those voters who vote in part because

symbolic reasons: Their votes for a marginal candidate *symbolises* the (unrealisable) victory of the policies they support. But if such symbolic voting for a marginal candidate reduces the tally of an intermediate candidate, it may diminish the chance of their preferred policies being implemented. Consider the citizens who voted for Ralph Nader in Florida in the 2000 U.S. Presidential Elections. Some of these voters may have voted for Nader because the action of voting for him symbolised the policies that they prefer on *e.g.* the environment. The instrumental function of the vote — selecting a president — makes the act of voting an endorsement of a candidate. The fact that voting constitutes an endorsement enables a symbolic transfer of value from the candidate and his or her policy platform to the act of voting for the candidate. This symbolic value gave some voters a reason to vote for a candidate like Ralph Nader even when he had no chance of winning. Now many of these voters presumably expected the policies of a Republican Bush administration to be much worse for the environment than those of a Democratic Gore administration. Given that George W. Bush carried the election with a whisker, we may assume that if the Nader supporters in Florida had voted for Gore instead, they would have enjoyed policies more closely aligned with their preferred ones. Was it rational for them to still vote for Nader rather engage in vote-trading with Gore supporters in other states? It was, if they did not expect the election to be as close as it turned out to be. But suppose that they did expect it to be that close. Then they would have attributed symbolic value to an action (voting for Nader) that they expected to undermine the ultimate values that were the source of the symbolic value (Nader's policies on, say, the environment). They would, in brief, have preferred an action that led to a less highly valued state of affairs than what was possible. For what reason? The only reason is that that action symbolises the more highly valued state of affairs. But this is not a rationally defensible reason when the action causes that very value to be less well realised. Valuing the vote in a way that makes the valued culmination outcome less likely, *for no good reason*, is inconsistent (by being self-defeating) and surely irrational.

But this is a rather special case. In general, it need not be true that the non-instrumental value of an action or a process derives from the outcome of the process itself. A given action could represent

they value voting not merely instrumentally. If this is the *only* way voting can ever be rational, then my argument is only strengthened by it.

something different from its own actual or typical causal outcome. An action could derive symbolic value from other intrinsically valued outcomes, from other instrumentally valued actions, or from other intrinsically valued actions. In particular, an action may have symbolic value because it instantiates a *principle*. Principles have the function of grouping actions together in classes. “By adopting a principle, we make one action stand for many others and thereby we change the utility or disutility of this particular action” (Nozick 1993, p. 18). Evidential or symbolic value, then, could be mediated by principles rather than derive directly from a valued or disvalued state of affairs. Thus certain actions or objects are valued because they express, are evidence of, or symbolise principles held dear by the decision-maker. The derivation of value from principles may happen at two levels. An action can “borrow” value from the other actions covered by the principle, and these actions may have instrumental value. This shows how it may be rational to care about what the classical utilitarians called the *tendency* of actions, as opposed to the actual consequences of a single action. It is only when a class of actions has been defined that it is possible to talk of the effects an action *tends* to have. A principle does precisely that: It identifies an action as belonging to a certain class of actions, namely those covered by the principle. This allows an individual action to derive symbolic value from the instrumental tendency of its class. At a more fundamental level, each action can derive symbolic value from the principle itself (when the assignment of certain actions to a class covered by a certain principle is intrinsically valued), and even from the very fact of acting on principle. When values take the form of principles, acting on those principles can be *ipso facto* valuable. Being a person who acts on principles is itself something we may value, and actions that conform with principles thus derive evidential value from that goal.

The possibility that actions symbolise other actions by instantiating principles solves the puzzle of derivative non-instrumental valuation. Principles can be principles about *how to do things*, about how to achieve (culmination) outcomes. An action covered by the principle could then have value that is derivative — because it would only be covered by the principle as an action with a certain aim, but not when considered *in vacuo* — and yet the value could be non-instrumental, because it would reflect the value of acting on the principle or symbolise the value of the other actions

covered by it, which could in general be different from the actual instrumental value of the specific action in question.

The inconsistency argument outlined in subsection 4.2 threatens the rationality of evidential or symbolic value directly derived from the represented states of affairs, but there is no inconsistency involved in evidential or symbolic value derived from principles.³³ An action can in this way take on symbolic value that withstands the possibility that in any specific instance the action has less highly valued consequences than do (some of) its alternatives, even when this is known with certainty. A *principled* pacifist, for example, may rationally prefer to participate in an anti-war demonstration rather than stay at home because that action expresses or symbolises his pacifist principle, even if he knows that the demonstration will make war more likely. He may rationally prefer this if his principles commit him to fight war through vocal participation in anti-war efforts (which perhaps *tend* to make wars less likely even if this particular demonstration has the opposite effect). A *non-principled* pacifist on the other hand — one who simply values the outcome in which there are as few wars as possible — cannot rationally have such a preference but must form an opinion on whether his participation in the anti-war demonstration on balance increases or diminishes the likelihood of war. Another example, drawn from history, is Josephus' account of the Jewish resistance at Masada in the AD70-73 uprising against Roman rule — according to the chronicler more than a thousand fighters preferred collective suicide with their families to being captured alive. Whether the account is accurate or not, the point is that people may rationally prefer to “die fighting” rather than *accepting defeat*, even when defeat is certain. Fighting to the end may symbolise victory in a way that capitulating does not.

In arguing that symbolic relations to a valued culmination outcome are not rationally sufficient for conferring value on actions above and beyond their causal efficacy, we pointed to the problematic possibility that an action that is highly symbolic of a good outcome may in some instances causally prevent it from being brought about. The property of symbolising a valued outcome is

³³Of course the principles themselves could be inconsistent, but this is not a problem specific to non-instrumental valuation of processes based on principles. Rationality demands that principled valuation be consistent, whether the object of the principles is culmination outcomes or actions.

not a reason to give preference to an action that undermines that very outcome. *Principle-based* process-dependent preferences may seem to encounter the same problem — they may also confer value on actions that lead to culmination outcomes that are less valuable than other outcomes that could have been brought about. In such cases, the rational evaluation of action must balance the value of achieving culmination outcomes in the best way against that of achieving the best culmination outcome. A poignant example of this type of conflict occurs when adherence to legality and due process in politics leads to a state in which they are both likely to be undermined, or when liberal tolerance benefits intolerant movements, as one might argue happened in Algeria and Venezuela in the 1990s. But hard cases are not irrationalities. There is nothing irrational in having competing concerns (acting on principle versus achieving valuable outcomes); whereas there is something irrational about letting an action derive value from an outcome it undermines.

Principles are different from other values in that they tend to be more profoundly bound up with our sense of who we are; they are what Bernard Williams call “commitments.” As such, one may change one’s “mere” preferences, but altering one’s commitments is a much more radical transformation, frequently accompanied by reflective “soul-searching” or at least a retrospective acknowledgment of having become “a different person.” On a formal level, this complexity has been analysed as the difference between first-order preferences and second-order preferences or meta-preferences (and even third-order preferences, about how to resolve conflicts between first- and second-order preferences). Value judgments that are instances of principles will reflect higher-order preferences, while direct value judgments that do not reflect commitments to principles are typically lower-order preferences. When I choose between actions *A* and *B*, I may first-order prefer the consequences of action *A* to those of *B*, but I may second-order prefer to prefer actions that have consequences like action *B*. (Note that this does not mean that I have a second-order preference for the *consequences* of *A*.) In this case, even if I do (first-order) prefer action *A* for instrumental reasons, action *B* may derive a symbolic value from the second-order preference which outweighs the first-order preference. Even if on the first order of preferences, it is not worthwhile for me to vote (rather than abstain, or engage in vote-trading), a second-order preference for

agency may mean I *prefer to prefer* actively participating in the vote. The second-order preference for agency, in general, may put positive value on my participation in bringing about the outcomes that matter to me, even if the value is not warranted by my participation's effect on the probability of success. This would account for the positive preference many people have for actively participating in the vote.

A thorough theory of rational preferences must have room for higher-order preferences, and it is not implausible to argue that it would be a less than fully rational value system that consisted only of first-order preferences. The imaginary agents with such simple preferences common in economic theory and other social sciences that have followed the discipline have been variously labelled “rational fools” (Sen 1976) or “wantons” (Hirschman 1982), and it has been argued that the conventional model with a single preference ordering is too impoverished to fully capture the value systems of real individuals. If preferences over comprehensive outcomes are rooted in *principled* procedural preferences, such value systems are perhaps more rational than those defined over culmination outcomes only.

It should be clear that this account of process-dependent preferences has very different implications from the indirect instrumentality account we analysed and rejected in the previous section. We concluded above that considerations of indirect instrumentality do not bind a society in the way they bind an individual, so that social evaluation may be permitted to ignore process-dependent preferences for the sake of better outcome achievement when process-dependence is rationalised by indirect instrumentality. But no analogous distinction exists between the individual and the social level that should make us discount preferences over comprehensive outcomes that reflect the symbolic value of principles. These values are as genuine as first-order values about culmination outcomes, and if anything, it may be value systems restricted to the latter that are less than fully rational, not value systems that include derivative non-instrumental valuation. Theories of social evaluation that make room for individual values, then, need to consider seriously the possibility of individual procedural preferences in a way that has not been commonly done. The analytical and moral difficulties are serious. While it may be hoped that procedural and outcome-oriented

preferences often point in the same direction, this is more than we may realistically expect. Some of the most challenging moral and political problems arise from conflicts between procedure and outcome. Affirmative action (where many prefer a neutral admissions process to a non-neutral one, but a more diverse outcome to a less diverse one) and domestic security policy (where many prefer strong protection of individual rights, but also prefer more security to less) are but two contemporary examples.

7 Consequentialism, instrumental rationality, and process-dependent preferences

I finish this essay with an aside about consequentialism and instrumental rationality. One may ask whether preferences that are irreducibly defined over comprehensive outcomes are nonconsequentialist. The answer is that they may be, but not necessarily. I follow Bernard Williams (1973) in distinguishing consequentialist from non-consequentialist valuations of actions according to how such theories understand the relationship between statements about the rightness of actions and the goodness of states of affairs. In a consequentialist theory, saying that action A has value implies that it leads to a valuable state X , and saying that A is best (most valued) implies that X is the most highly valued state available to the agent. Preferences are defined over comprehensive outcomes when “ A has value” implies that A leads to a valued state X in the sense that X *consists* (in part) of action A being performed.³⁴

Non-consequentialist theories of action, on the other hand, do not posit such a relationship between statements about the value of actions and the value of outcomes. This can be either because they do not rank actions and outcomes by their goodness at all, or because they allow that an ac-

³⁴Williams suggests that it would not be compatible with consequentialism to admit that the valuable state X may consist of the action A being performed by a specific agent. Many authors seem to think that consequentialism cannot be agent-specific. Agent-specificity holds when knowing who performs an action is essential in order to evaluate a situation. Perhaps the aversion to agent-specificity comes from the intuition that a moral theory must be impartial; that is, the name of the agent should not matter. But agent-specificity does not violate that. It could be characterised in entirely general terms, focusing for example on the social roles of the agent. There is nothing inherent in the logic of consequentialism which rules out agent-specificity of this type.

tion can be right even if the comprehensive outcome it leads to is not the most highly valued of the achievable comprehensive outcomes. (Even consequentialist theories may allow that an action is best even if it does not lead to the most highly valued *culmination* outcome achievable, precisely because comprehensive outcomes may be valued differently from the culmination outcomes.) For example, a non-consequentialist theory, if it does allow value rankings of outcomes, may hold it wrong to break a promise even when the resulting outcome, including the fact that it was reached through the breaking of a promise, is the best possible comprehensive outcome all things considered. As a special case, a non-consequentialist theory may hold it wrong to break a promise even when the resulting comprehensive outcome is the best possible precisely because it leads to highest possible frequency of promise-keeping in the future.

On the other hand, I argued that preferences over comprehensive outcomes, if rational, are not instrumentally rational in a pure sense. This is not simply saying that such value systems must value *some* things intrinsically — any theory of instrumental rationality must allow for *that*, if instrumentality is to be an interesting concept. But we have argued that the value of actions and processes is not always intrinsic. When they have no instrumental connection with valuable culmination outcomes, they are sometimes worthless or even undefined, yet when they do have such a connection, their value is not reducible to their instrumental efficacy. We argued that this is because what is valued is their evidential or symbolic connection to other actions, and more generally, to principles to which the agent has a commitment. In other words, such actions are rational because they have derivative (evidential or symbolic) value that may be contingent on their instrumental function, but they are not *instrumentally* rational because some other available action may be the instrumentally most valuable one. Of course an instrumentalist may try to rescue an instrumental view by saying that the actions are valued just because they are instrumental in bringing about these evidential or symbolic connections. But note that even on that view, the actions are valued for being instrumental in placing themselves in a non-instrumental relation to other actions (the symbolic linking together of actions in a class through principles). That, as Nozick (1993) says, is for all intents and purposes non-instrumentality.

8 Conclusion

I started this essay by pointing to the lack of consensus on the question of whether it is rational to have preferences over comprehensive outcomes. The answer is important because it deeply affects the way in which individual values need to be taken into account in collective decisions. Preferences in the sense of considered value judgment can be deemed irrational if they entail inconsistencies that cannot be sustained in reflective equilibrium. I identified *process-dependence* and *non-separable process-dependence* as potential sources of such inconsistencies. One argument why process-dependence is irrational is that it is like caring about sunk costs, but I demonstrated that this argument rests on a confusion. Still, preferences over comprehensive outcomes sometimes have the puzzling feature that they involve non-instrumental yet derivative valuation of actions. The second argument for irrationality holds that this makes such value systems directly self-defeating and therefore irrational. I proposed two accounts of why they can nevertheless be rational. The first account is based on the indirect instrumental value of directly non-instrumental valuation. Because of imperceptible effects of single actions, it may be that only a value system which is directly self-defeating can avoid being indirectly self-defeating, and the latter may be a more serious obstacle to the achievement of valued outcomes than the former. This account, if right, entails that only preferences over culmination outcomes need be taken fully into account in collective decision-making. I argued, however, that it is not a satisfactory account, because it misinterprets non-instrumental valuation of actions as a *solution* to the problem of imperceptible individual effects. That is not how these values are understood by those who hold them. Instead, actions are valued non-instrumentally because there are other forms of derivative value than merely instrumental value. Actions can derive value through evidential and symbolic connections, and in particular by instantiating *principles* (the evidential or symbolic connections may be contingent on an instrumental connection without being reducible to it). Principled value systems are no less rational (and may be more rational) than value systems with only first-order preferences. I conclude that preferences over comprehensive outcomes can indeed be rational. The consequences of this conclusion are far-reaching: It implies that a whole range of normative theories of social

choice, and most importantly the reigning versions of normative economics, systematically involve an arbitrary rejection of a large part of people's preferences.

References

- Andreoni, James and John Miller. 2002. "Giving According to Garp: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica* 70(2):737–753.
- Blount, Sally. 1995. "When Social Outcomes Aren't Fair: The Effect of Causal Attribution on Preferences." *Organizational Behavior and Human Decision Processes* 63:131–144.
- Brandts, Jordi and Gary Charness. 1999. "Retribution in a Cheap-Talk Experiment." Universitat Pompeu Fabra Working Paper.
- Broome, John. 1999. *Ethics out of Economics*. Cambridge, UK: Cambridge University Press.
- Festinger, Leon. 1962. "Cognitive Dissonance." *Scientific American* 207(4):93–107.
- Festinger, Leon and James M. Carlsmith. 1959. "Cognitive Consequences of Forced Compliance." *The Journal of Abnormal and Social Psychology* 58:203–210.
- Hirschman, Albert O. 1982. *Shifting Involvements*. 20th anniversary edition ed. Princeton, NJ: Princeton University Press.
- Kahneman, Daniel, Jack L. Knetsch and Richard Thaler. 1991. "The Endowment Effect, Loss Aversion, and the Status Quo Bias." *Journal of Economic Perspectives* 5(1):193–206.
- Kahneman, Daniel and Richard Thaler. 1991. "Economic Analysis and the Psychology of Utility: Applications to Compensation Policy." *The American Economic Review* 81(2: Papers and Proceedings of the Hundred and Third Annual Meeting of the American Economic Association):341–346.
- Lind, E. Allan and Tom R. Tyler. 1988. *The social psychology of procedural justice*. New York: Plenum Press.
- Lyons, David. 1965. *Forms and Limits of Utilitarianism*. Oxford: The Clarendon Press.
- Nozick, Robert. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

- Nozick, Robert. 1993. *The Nature of Rationality*. Princeton: Princeton University Press.
- Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Samuelson, Paul A. 1938. "A Note on the Pure Theory of Consumers' Behaviour." *Economica* 5:61–71.
- Schuessler, Alexander A. 2000. *A Logic of Expressive Action*. Princeton, NJ: Princeton University Press.
- Sen, Amartya. 1973. "Behaviour and the Concept of Preference." *Economica* 40:241–259.
- Sen, Amartya. 1976. "Rational Fools: A Critique of the Behavioural Foundations of Economic Theory." *Philosophy and Public Affairs* 6(4):318–344.
- Sen, Amartya. 1987. *On Ethics and Economics*. Oxford and Cambridge, MA: Blackwell.
- Sen, Amartya. 1993. "Internal Consistency of Choice." *Econometrica* 61(3):495–521.
- Sen, Amartya. 1995. "Rationality and Social Choice." *The American Economic Review* 85(1):1–24.
- Sen, Amartya. 1997. "Maximization and the Act of Choice." *Econometrica* 65(4):745–779.
- Sen, Amartya. 1999. *Development as Freedom*. New York: Anchor Books.
- Williams, Bernard. 1973. A critique of utilitarianism. In *Utilitarianism: for and against*, ed. J. C. C. Smart and Bernard Williams. Cambridge, UK: Cambridge University Press pp. 75–155.