Hume's experimental psychology and the idea of erroneous preferences

Robert Sugden

School of Economics, University of East Anglia, Norwich NR4 7TJ, UK

r.sugden@uea.ac.uk

15 April 2020

Abstract: Hume's *Treatise of Human Nature* is not only a canonical text of philosophy, but also a pioneering work of psychology, anticipating many findings of modern behavioural economics. According to Hume's theory of mind, the concept of rationality does not apply to choices or moral judgements. But in in his theory of justice, Hume explains preference reversals between smaller-sooner and larger-later options in terms of far-sighted 'true' preferences and psychologically-induced errors of short-sightedness. Anticipating a common idea in behavioural welfare economics, he proposes a role for government in helping individuals to overcome self-control problems in acting justly. I examine Hume's position and assess its coherence. I conclude that Hume's theory of mind is consistent and psychologically well-grounded, but does not support the concepts of true preference and error that appear in his theory of justice. However, the fundamental logic of that theory does not depend on assumptions about self-control problems.

Acknowledgements: An earlier version of this paper was presented at a conference on 'David Hume, economic rationality, and policy' at New York University. I thank participants at that conference and an anonymous reviewer for insightful comments. This research was supported by the Economic and Social Research Council of the UK (award no. ES/P008976/1) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 670103.

JEL classifications: B12 (history of economic thought: classical); D91 (role and effects of psychological, emotional, social and cognitive factors on decision making)

Keywords: Hume; *Treatise of Human Nature*; experimental psychology; true preference; self-control

Declarations of interest: none

Hume's experimental psychology and the idea of erroneous preferences

David Hume's *Treatise of Human Nature* (1739-40/ 1978) has usually been read as a canonical work of philosophy, as that discipline is now understood.[1] Unquestionably, it has made lasting contributions to moral philosophy, political philosophy, philosophy of mind and philosophy of science. Its analysis of conventions is now also widely viewed as a major contribution to game theory (Lewis, 1960; Sugden, 1986; Binmore, 1994, 1998). It is less well recognised that the *Treatise* proposes a theory of human psychology that anticipates many findings of modern experimental research. What is particularly interesting to me as an economist, the *Treatise* anticipates findings that are now seen as fundamental to behavioural economics.

My interest in this is not merely historical. When what is now called 'behavioural economics' began to gain attention in the 1980s, most economics was based on the neoclassical model of the individual as a rational agent. Initially, the rhetoric of behavioural economics was one of opposition to the neoclassical approach, but from the outset, rational choice models were used as templates. Many of the most influential theories in behavioural economics – for example, prospect theory (Kahneman and Tversky, 1979), the reciprocity-based theory of fairness (Rabin, 1993), and the theory of inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) – have been constructed by revising or adding to pre-existing models of utility maximisation to accommodate psychological insights. In contrast, Hume was trying to make sense of his psychological discoveries at a time when almost all the tenets of rational choice theory were unknown. It is not obvious that this was to his disadvantage. The premise of my paper is that, because Hume did not think about human psychology in relation to the theory of rational choice, he may have recognised properties of that psychology, and implications of those properties for economic and social life, that modern behavioural economists have failed to see.

In a previous paper, I have argued that the *Treatise* contains the elements of a theory of human decision-making that is radically different from modern decision theory, and more

[1] Intentionally, my paper is about Hume's *Treatise*, not the totality of Hume's views on the topics he addresses in that book. Hume 'cast anew' the arguments of the *Treatise* in his later *Enquiries*, asking that the latter be treated as the definitive statement of his philosophical principles (1748-51/1975: 2). Nevertheless, I share the view of many readers that some of the deepest and most original ideas of the *Treatise* have been edited out of the *Enquiries*. In particular, many of Hume's psychological hypotheses and discoveries are lost in this editing. Whenever I refer to what Hume thought, I mean what he thought when writing the *Treatise*. Unless otherwise stated, all subsequent page references are to Hume (1739-40/ 1978).

firmly grounded in empirical psychology (Sugden, 2006). In that paper, I discuss some modern philosophers who, reading the *Treatise* through the lens of rational choice theory, have thought Hume's psychological arguments confused, or valid only on naïve or antiquated assumptions. I try to show that those arguments are part of a coherent attempt to build a theory of human nature that is consistent with scientific observation. In the present paper, I focus on what the *Treatise* has to say about normative issues, viewed in relation to a significant problem that behavioural economists continue to struggle with.

The problem is that standard methods of normative economics use preference-satisfaction as an evaluative criterion, and depend on the assumption that individuals have well-defined preferences that are revealed in their choice-making behaviour; but these assumptions are called into question by the findings of behavioural economics. In particular, experimental evidence shows that people's revealed preferences are highly *context-dependent* – that is, sensitive to contextual features of decision environments that cannot credibly be claimed to be relevant to decision-makers' interests or well-being.

One of the most common ways of dealing with this problem is to model human behaviour as resulting from some combination of *latent* (or 'true' or 'underlying') preference and error. I will say that a person's preferences over some domain of possible objects of choice are *integrated* if they are complete, transitive and context-independent. If a psychologically plausible model, combining integrated latent preferences with an error mechanism, could be constructed and fitted to choice data, we would have an operational method for reconstructing those preferences; behaviour that was inconsistent with latent preferences could be attributed to error. Variants of this approach have been proposed or endorsed by, among others, Bleichrodt et al. (2001), Camerer et al. (2003), Sunstein and Thaler (2003), Kőszegi and Rabin (2007), Bershears et al. (2008), Salant and Rubinstein (2008), Thaler and Sunstein (2008), Manzini and Mariotti (2012), and Le Grand and New (2015).[2]

As argued by Infante et al. (2016), the various form of this approach share a common feature: in one way or another, they use a model of an *inner rational agent*. Less

[2] A related approach of *behavioural revealed preference*, proposed by Bernheim and Rangel (2009) and Bernheim (2016) identifies and corrects supposed errors in a slightly different way. This approach does not invoke any concept of latent preference, accepting as data only preferences that have actually been revealed in an individual's choice behaviour. However, data from situations in which the decision-maker 'incorrectly perceives the choice set' are ignored (Bernheim and Rangel, 2009: 83).

metaphorically, human decision-making is modelled as if its psychological substrate consisted of a sub-routine of rational choice, linked to the world of actual choice problems by error-prone mechanisms of perception and volition. But this model lacks credible psychological foundations. On an uncharitable interpretation, this approach amounts to an attempt to conserve the theory of rational choice in the face of disconfirming evidence by reclassifying that theory's prediction errors as errors on the part of the individuals whose behaviour the theory is supposed to explain (Infante et al.: 23, note 5; Berg and Gigerenzer, 2010).

If one entertains the possibility that the use of inner rational agent models is a false move by behavioural economics, it is of interest to ask how Hume dealt with normative issues, given that he presented his psychological theory as an explanation of context-dependent desires and beliefs. Did he too use concepts of true preferences and error? If not, how was he able to make normative statements about social institutions? A short answer to the first question is that Hume uses these concepts very rarely – but not quite never. Hume's discussion of justice includes some significant passages which, for a modern reader, seem to anticipate the latent preference approach, as that is now used in behavioural economics. My aim in this paper is to understand Hume's position and to assess its coherence.

1. Hume's experimental methodology

Hume is explicit in presenting the *Treatise* as a work of experimental psychology. Its full title is: *A Treatise of Human Nature: being an attempt to introduce the experimental method of reasoning into moral subjects*. The objective of the book is set out in its Introduction. It is to be a contribution to the 'science of man' – a subject that is fundamental to all the other sciences. Hume's guiding idea is that everything that we count as knowledge or understanding, whatever reference it may make to the world outside our minds, ultimately exists in the form of human mental states. All scientific reasoning is ultimately a matter of operations of the human mind. Thus, we need scientific explanations of how, as a matter of empirical psychology, we arrive at the mental experiences of knowledge and understanding. Hume's intention is to offer such explanations.

His methodology is to be that of the natural sciences: 'And as the science of man is the only solid foundation for the other sciences, so the only solid foundation we can give to this science itself must be laid on experience and observation' (xvi). And then:

> For to me it seems evident, that the essence of the mind being equally unknown to us with that of external bodies, it must be equally impossible to form any notion of its powers and qualities otherwise than from careful and exact experiments, and the observation of those particular effects, which result from its different circumstances and situations. (p. xvii)

Hume accepts that the science of man cannot use the kinds of controlled experiments that are used in the natural sciences. In a move that might surprise modern experimental psychologists, his analogue of conducting a physics experiment is an experiment on his own mind – 'placing myself in the same case with that which I consider'.[3] He argues that this method would be unsatisfactory:

> 'tis evident this reflection and premeditation would so disturb the operation of my natural principles, as must render it impossible to form any just conclusion from the phænomenon. We must therefore glean up our experiments in this science from a cautious observation of human life, and take them as they appear in the common course of the world, by men's behaviour in company, in affairs, and in their pleasures. (p. xix)

This method is not 'experimental' in a common modern sense of the word, in which an experiment is a manipulation of natural objects in an artificially created environment. But Hume's method of enquiry is not as different from modern empirical science – and in particular, from modern psychology – as the previous quotation might suggest.[4] It is true that many of Hume's 'experiments' *are* observations of ordinary human life. One of his most common strategies is to point to regularities in human behaviour whose existence the reader will immediately recognise, but from which he draws startling conclusions. These observations are psychological analogues of the observation of the darkness of the night sky and the inference that (contrary to what we might naturally have thought) we do not live in an infinite, homogeneous and static universe. Commonplace observations, properly understood,

---

[3] Or perhaps they would not be surprised. Describing his early collaboration with Amos Tversky, Daniel Kahneman (2011: 4–10) says that their research routine consisted of regular conversations, often during long walks: '[W]e invented questions and jointly examined our intuitive answers. Each question was a small experiment… We believed – correctly, as it happened – that any intuition that the two of us shared would be shared by many other people as well, and that it would be easy to demonstrate its effects on judgments.'

[4] Demeter (2012) discusses Hume's method in the context of eighteenth century science. He concludes, I think rightly, that its *processing* of empirical data is experimental, but its *production* of these data is not. But I think Demeter underestimates the power of this method when he argues that without 'contrived experience' (i.e. experimental data) 'asking specific questions about the reliability of a theory of human nature is hardly possible' (p. 582).

can confirm or disconfirm received theories. But the *Treatise* is also full of experimental designs that the reader is invited to carry out for herself, using herself as the subject and recording her own psychological responses to some real or imagined external stimulus. Hume tells the reader what she will find. How does he know? Presumably because he has carried out the same experiments on himself. It requires a special kind of self-awareness to make unselfconscious responses to imagined stimuli while simultaneously keeping track of your states of mind. Judging how likely it is that other people's mental responses will be similar to yours requires some understanding of the general features of human psychology, combined with an ability to sense what kinds of operations your mind is performing. These are skills that a good psychologist needs, and Hume is a master of them. For all its informality, this is genuine experimental science.

2. Hume's theory of mind

Hume's account of human psychology is fundamentally different from those that are implicit in modern models of rational choice – in which I include both the formal models used in neoclassical economics and game theory, and the informal models used in analytical philosophy.[5] The idea that individuals have integrated preferences – whether preferences that are directly revealed in decisions, or latent preferences that combine with error-generating mechanisms – does not fit easily (if at all) into Hume's framework.

Hume develops a theory of mental states, or 'perceptions of the human mind'. Perceptions are divided into *impressions* and *ideas*. In essence, the distinction between impressions and ideas is the distinction between feeling and thinking. Ideas are the 'faint images [of impressions] in thinking and reasoning' (pp. 1–2). *Beliefs* are ideas that are associated with a particular kind of feeling, which Hume calls 'force' or 'vivacity' (pp. 94–98, 629). Impressions are subdivided into *sensations* (or *original impressions*) and *reflections* (or *secondary impressions*). Sensations, for example of heat and cold or hunger and thirst, reach the mind from the sensory mechanisms of the body; they are the origins of all other perceptions. Reflections ('*viz.* passions, desires, and emotions') are affective states which (in modern language) have positive or negative valence (pp. 7–8, 275–276). *Desire* is the mental state in Hume's system that is closest to the modern concept of preference. It is an 'emotion of propensity' which 'unites us to' the idea of some object (414, 439). That is, it is a passion that focuses on the idea of some object and induces us to approach, possess or

[5] This section of the current paper draws on Sugden (2006).

consume it. The Humean mental state that is closest to choice is *volition*. This is the felt experience of intentional action, 'the internal impression we feel and are conscious of, when we knowingly give rise to any new motion of our body, or new perception of our mind' (399). Hume's theory is about how the mind processes sensations, ideas and reflections to create further ideas, further reflections, and volitions. Three fundamental properties of this theory are particularly important for my arguments in this paper.

First, Hume's theory is *non-propositional*. Hume does not assume, as many analytical philosophers do, that language is prior to thought. In communicating with his readers through the medium of print, he has no option but to try to describe ideas and impressions in words, but he is conscious that any such translation must be inadequate:

> 'tis very difficult to talk of the operations of the mind with perfect propriety and exactness; because common language has seldom made any very nice distinctions among them, but has generally call'd by the same term all such as nearly resemble one another. (p. 105)

Thus, there are distinctions between feelings 'of which 'tis impossible to give any definition or description, but which everyone sufficiently understands' (p. 106). Hume sees it as a 'most fertile source of error' that metaphysicians 'use words for ideas, and ... talk instead of thinking in their reasonings' (pp. 61–62). This property of Hume's theory must be kept in mind in any attempt to translate it into the language of rational choice theory. Rational choice theory is based on axioms of consistency that are imposed on preferences and, in Savage's (1954) canonical treatment, thereby on beliefs. If preferences are represented as propositions, or as attitudes to propositions, properties of consistency among preferences can be construed as analogous with principles of propositional logic. Having preferences that contravene such properties can then be characterised as analogous with making errors of logic. This is true even if, as in many versions of rational choice theory, preferences are interpreted as propositions about subjective attitudes. If, in contrast, preferences are understood as non-propositional feelings, it is possible to claim that certain combinations of such feelings are psychologically dissonant, and perhaps even to call this kind of dissonance 'inconsistency'; but it is not clear what it would mean to say that a person with dissonant feelings has made an error.

Second, Hume insists that neither passions nor volitions can be called reasonable or unreasonable:

> Reason is the discovery of truth or falshood. Truth or falshood consists in an agreement or disagreement either to the *real* relations of ideas, or to *real* existence and matter of fact. Whatever, therefore, is not susceptible of this agreement or disagreement, is incapable of being true or false, and can never be an object of our reason. Now 'tis evident our passions, volitions, and actions, are not susceptible of any such agreement or disagreement; being original facts and realities, compleat in themselves, and implying no reference to other passions, volitions, and actions. 'Tis impossible, therefore, they can be pronounced either true or false, and be either contrary or conformable to reason. (p. 458)

A belief about a matter of fact – either about the existence of some object, or about a 'connexion of causes and effects' – can be called reasonable or unreasonable, by virtue of its correspondence or lack of correspondence with the facts it supposedly represents. But the passions and volitions that arise in response to such beliefs do not represent anything; they are psychological facts in their own right. Even if the belief that prompts an action is erroneous, it is only 'in a figurative and improper way of speaking' that the action itself can be called unreasonable (p. 459). Hume is explicitly rejecting the idea that there can be a theory of rational choice.

Third, Hume's theory is *dynamic*. It is a theory of how mental states come into and go out of existence, the temporary existence of one mental state causing the temporary existence of another. For Hume, instability of mental states is a fundamental property of human psychology: ''Tis impossible for the mind to fix itself steadily upon one idea for any considerable time; nor can it by its utmost efforts ever arrive at such a constancy' (p. 283). This is quite unlike rational choice theory, in which the content of the mind is implicitly modelled as a stock of preferences and beliefs that are in some sort of equilibrium with one another. Facing any particular choice problem, an agent in rational choice theory consults this stock and reads off those that are relevant to the case in hand. In Hume's theory, in contrast, mental operations are understood as *transitions* between one mental state and another.

The transition mechanism is one of *association*. The existence of one mental state facilitates the creation of other mental states that are linked to the original one by salient relationships of particular kinds. Transitions between ideas can be induced by relations of resemblance, contiguity, or cause and effect. Transitions between impressions are induced only by resemblance, which Hume seems to interpret in terms of affective valence: positive emotions induce positive emotions, and negative emotions induce negative emotions.

Impressions can give rise to associated ideas, and ideas can give rise to associated impressions. Hume gives particular attention to *double relations* in which associations of impressions and associations of ideas are mutually reinforcing (pp. 283–284, 288–289).

Mental associations can produce patterns of thought which, if viewed in the propositional framework of rational choice theory, would seem clearly irrational. Hume's opening example of a double relation is about *pride*. On Hume's analysis, pride is a positive feeling about yourself, caused by thinking favourably about objects that are related to you. As a matter of psychological fact, just about any object, any favourable characteristic, and any kind of relationship can cause pride. Thus: 'Men are also vain of the temperature of the climate, in which they were born' (pp. 276–279; 306–307). Here is how the mechanism might work. Suppose I am enjoying an English summer day of sunshine and showers. I read in my newspaper that there is extreme heat in southern Europe. I feel a sense of pleasure about the temperateness of the English climate. This is an emotion with positive valence. I think about the fact that I was born in England. This is a relation of contiguity between me and the English climate. Pride is an emotion about myself, and it too has positive valence. The double relation of contiguity and valence facilitates a transition to a feeling of pride about the English climate. If it were meaningful to ask whether this feeling is reasonable, it seems that the answer would have to be 'No'. (How can I take credit for where I was born? All things considered, is the English climate really better than that of southern Europe?) But that does not alter the psychological facts about what my feelings are on the day, and how they were caused.

As supporting evidence for his theory, Hume analyses many cases in which beliefs and emotions are affected by contextual cues that are psychologically salient but (given the questionable premise that rational appraisal is possible at all) rationally irrelevant. Here are just a few examples, selected for their connections with behavioural economics.

The strength and 'violence' of a person's desire for any given object is affected by its closeness to him in time, space and imagination. Thus, according to Hume, if you want 'to govern a man, and push him to [choosing some object]', the best method is 'to place the object in such particular situations as are proper to increase the violence of the passion' (p. 419). This principle underlies many modern 'behavioural insights', exemplified by Sunstein and Thaler's (2003: 1164) 'libertarian paternalist' cafeteria director, who places the healthier options in more prominent positions on the counter. A related effect, which might also be exploited by the cafeteria director, is that a person's appetite for a dish – her desire for the

experience of eating it – is affected by the beauty of the table setting (pp. 394–395). This is one of Hume's double relations: the table setting is associated with the dish by contiguity, and desire and the sense of beauty both have positive valence. Hume recognises the endowment effect: 'Men generally fix their affections more on what they are possess'd of, than on what they never enjoy'd' (p. 482). He does not offer an explanation of this effect, but it might be analysed as another double relation, similar to pride. My possession of a desirable object establishes a mental association between the idea of the object and the idea of myself; since desire and love of myself share positive valence, self-love increases the intensity of my desire for what I already possess. Hume also recognises the tendency for preferences between (what are now called) 'smaller sooner' and 'larger later' options to reverse as the time of 'sooner' approaches, explaining this as an effect of contiguity on desire (pp. 536–537). Hume's analysis of this tendency is crucial for the topic of this paper; I will examine it more closely in Section 5.

As I noted in the introduction, behavioural economists often explain context-dependent preferences as resulting from errors, and define errors in relation to some concept of latent preference. My objective in this paper is to ask whether Hume's theory of mind allows any analogous concept. On the face of it, the answer is 'No'. If preferences are comparative desires, they are original facts and realities, incapable of being pronounced true or false. Having arrived at empirical explanations of people's actual feelings of desire, there seems to be nothing left for a Humean theorist to do: there is no space in the theory for a concept of error. But is this conclusion too quick? In the next section, I explore some of what might seem to be ways of escaping it.

3. Can there be a Humean concept of latent preference?

3.1 True and false beliefs

One potential escape route is to exploit Hume's equivocation about whether actions can be unreasonable. Recall his concession that if the mental process that produces an action is affected by a false judgement, there is a 'figurative and improper' sense in which that action is unreasonable. So is the issue merely one of semantics? Can we not choose to use the word 'unreasonable' in this other sense?[6]

[6] To do this would be to use the concept of error as it is used in the behavioural revealed preference approach of Bernheim and Rangel, mentioned in footnote 2.

There are two related problems with this escape route. The first is that it can deal only with cases in which the context-dependence of revealed preferences really is the result of false beliefs; but if Hume's theory of mind is correct, context-dependence is a much more general phenomenon. Context-dependence is generated by the mind's responses to relations of resemblance, contiguity, and cause and effect. In most of Hume's examples, the perception of these relations does not involve error. Take the case of pride. In my version of Hume's example, my enjoyment of the English summer day is just a feeling; it is neither true nor false. My belief that there is extreme heat in southern Europe is true. The association between me and the English climate is my belief that I was born in England, and that is true too. If my pride is unreasonable, that has nothing to do with false beliefs.

The second problem is that, if transitions between mental states depend on subjective relations of resemblance and contiguity, the question of whether a volition was caused by a false belief may not have a clear answer. Consider one of Hume's more graphic examples of how contiguity affects beliefs:

> Thus a drunkard, who has seen his companion die of a debauch, is struck with that instance for some time, and dreads a like accident for himself: But as the memory of it decays away by degrees, his former security returns, and the danger seems less certain and real. (p. 144)

Toning down this case, suppose I customarily drink rather more alcohol than the UK Chief Medical Officers' well-known recommended limits. I am aware that this carries some health risk, but I have no idea what this risk means in terms of any medical statistics. Enjoying my visits to my local pub, I judge the risk to be worth taking. One day I learn that one of the pub's regular customers, a man who drank much more than I do, has died of liver disease. For some time, I reduce the amount I drink. But as the memory of his death recedes – and still with no more knowledge of the medical statistics – I revert to my previous habits. It is clear enough that the mechanism proposed by Hume is at work. But was one of my levels of drinking caused by a false belief, and if so, which? The reality is that my perceptions of risk are non-propositional: they are feelings – in Hume's language, feelings of vivacity that are associated with ideas such as that of my premature death – which fluctuate according to the cues to which I am exposed. My volitions are caused by those fluctuating perceptions, not by beliefs that can be expressed in terms of probabilities that might be true or false.

3.2 Pleasure and pain

Hume occasionally uses expressions which, to a modern reader, might suggest a position similar to that of classical utilitarianism – the idea that individuals seek to maximise the net balance of pleasure over pain. For example (in a passage that is immediately followed by a definition of 'good' and 'evil' as synonyms of 'pleasure' and 'pain'): 'DESIRE arises from good consider'd simply, and AVERSION is deriv'd from evil. The WILL exerts itself, when either the good or the absence of the evil may be attain'd by any action of the mind or body.' (p. 439). And:

> The chief spring or actuating principle of the human mind is pleasure or pain; and when these sensations are remov'd, both from our thought and feeling, we are, in a great measure, incapable of passion or action, of desire or volition. (p. 574)

Could erroneous choices be defined as ones that fail to maximise net pleasure?

I think not, for three reasons. First, Hume is not proposing a theory of rational choice in which the desire to perform an action is induced by the belief that that action will bring about pleasurable consequences. His hypothesis is non-propositional: the *idea* of pleasure tends to induce the *feeling* of desire, which then activates the will. Second, Hume recognises that some desires and volitions are *not* activated by ideas of pleasure and pain:

> Beside good and evil, or in other words, pain and pleasure, the direct passions frequently arise from a natural impulse or instinct, which is perfectly unaccountable.[7] Of this kind is the desire of punishment to our enemies, and of happiness to our friends; hunger, lust, and a few other bodily appetites. These passions, properly speaking, produce good and evil and proceed not from them, like the other affections. (p. 439)

Here Hume seems to be recognising the distinction between *wanting* and *liking*.[8] With his usual psychological insight, he is proposing that (for example) the feeling of hunger and the associated desire to eat are activated directly by the body's need for nutrition, and not by a desire for the pleasure of eating. The pleasure of eating is merely the sensation of satisfying that desire. Third, Hume recognises that pleasure is not a single type of feeling: ''tis evident, that under the term *pleasure*, we comprehend sensations, which are very different from each other, and which have only such a distant resemblance, as is requisite to make them be

[7] I take in that when Hume says that an impulse is 'unaccountable', he means that it comes from outside the mind. The kinds of impulses he has in mind have biological functions – roughly, human survival and reproduction – which he recognises and attributes to 'Nature' (e.g., pp. 119, 417).

[8] This distinction is important in modern neuroscience, particularly in relation to drug addiction. See, for example, Robinson and Berridge (1993).

express'd by the same abstract term'. He gives the example of the difference between the pleasure of music and the pleasure of wine (p. 472). Hume's non-propositional theory of mind gives us no reason to suppose that different types of pleasure and pain can be integrated into a one-dimensional maximand.

3.3 Calm and violent passions

Hume differentiates between *calm* and *violent* passions:

> Now 'tis certain, there are certain calm desires and tendencies, which, tho' they be real passions, produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation. These desires are of two kinds; either certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children; or the general appetite to good, and aversion to evil, consider'd merely as such. When any of these passions are calm, and cause no disorder in the soul, they are very readily taken for the determinations of reason, and are suppos'd to proceed from the same faculty, with that, which judges of truth and falshood. (p. 417)

Hume contrasts these calm passions with 'certain violent emotions of the same kind', such as the violent passion of resentment that a person feels immediately after someone has injured him. Both types of passion have effects on volition (pp. 417–418).

It is easy to see why Hume needs the concept of calm passions. There are many desires that are important in determining our day-to-day actions, but which we are not normally conscious of as desires. These are passions in Hume's technical sense, but not in the ordinary sense of the word. Hume wants to head off the thought that such desires are the product of reason. But, as in the case of false beliefs, might one not instead use different definitions? Why not define calm desires as reasonable, and say that actions that conflict with calm desires are the result of error?

The problem is that, as Hume says, violent emotions are *of the same kind as* calm ones. Although calm desires might be less volatile than violent ones, they are governed by the same mechanisms of association. Hume's theory gives us no reason to expect calm desires to be context-independent. For example, in going to work each morning, I might act on a calm desire for income. If I find my work fulfilling, I might also act on a calm desire to do my job well. (Alternatively, as Hume seems to suggest is possible, I might act on a calm sense of resentment towards an exploitative employer by being deliberately unproductive.) In caring for my children when I am not at work, I might act on a calm desire for their

welfare. These desires will sometimes come into conflict. It is a natural implication of Hume's theory that the relative influence on my actions of different calm desires will vary according to my current situation and the associations of ideas and impressions that it induces. Hume's conceptual framework gives us no way of characterising some situations as producing the correct weightings of those desires and others as producing errors.

3.4. Conventions of taste

The context-dependence of judgements is a potential obstacle to communication between people who are subject to different psychological cues. Hume introduces this problem and its solution in relation to judgements about the praiseworthiness of people's characters:

> 'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view. In order, therefore, to prevent those continual contradictions, and arrive at a more stable judgment of things, we fix on some *steady* and *general* points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation. (pp. 581–582)

Hume draws an analogy with judgements about beauty:

> In like manner, external beauty is determin'd merely by pleasure; and 'tis evident, a beautiful countenance cannot give so much pleasure, when seen at the distance of twenty paces, as when it is brought nearer us. We say not, however, that it appears to us less beautiful: Because we know what effect it will have in such a position, and by that reflexion we correct its momentary appearance. (p. 582)

The idea is that the language of judgement is based on conventions that fix the point of view from which judgements of particular types are to be made. For me to say, at a particular time and place, that Anna is more beautiful than Beatrice is not to say that, at that moment, I have a more vivid mental impression of Anna's beauty than of Beatrice's. I am expected to correct for context-dependent cues such as that I am standing closer to Anna, or the double relation that Anna is my daughter (which might work like pride: my love of myself gives vivacity to my perception of my daughter as beautiful). Perhaps I am also expected to correct for idiosyncrasies of personal taste. (I might say that Carla is more beautiful than Daphne *according to conventional tastes in beauty*, even though *to my eyes* Daphne is the more beautiful.)

This theory of judgement is used repeatedly by Hume. Garrett (2015: 117–145) identifies a common four-stage structure in Hume's analyses of *sense-based concepts*, such as those of beauty, virtue, probability and causation. First, for individuals separately, a

certain type of stimulus repeatedly activates a characteristic mental response, such as the pleasurable sensation that is activated by seeing certain types of human face. Second, individuals separately recognise this correspondence between stimulus and response and generalise it as an as abstract idea, such as 'a beautiful face'. Third, in interacting with one another, individuals converge on a conventional standard of judgement for that abstract idea; this standard is given by the characteristic mental response of an observer who is taking a certain kind of 'steady and general' point of view. Finally, that standard becomes accepted as the meaning of the abstract concept. Thus, although the sensations on which the concept is based are original facts and realities, judgements about the application of that concept can be correct or incorrect.

Crucially, however, the social significance of judgements of taste does not displace the force of private passions, desires and emotions. Thus, for Hume:

> The intercourse of sentiments, therefore, in society and conversation, makes us form some general inalterable standard, by which we may approve or disapprove of characters and manners. And tho' the *heart* does not always take part with those general notions, or regulate its love and hatred by them, yet they are sufficient for discourse. (p. 603)

And:

> A house may displease me by being ill-contriv'd for the convenience of the owner; and yet I may refuse to give a shilling towards the rebuilding of it. Sentiments must touch the heart, to make them controul our passions: But they need not extend beyond the imagination, to make them influence our taste. (p. 586)

In matters of private choice, actual desires are often more important than convention-based judgements of taste – and, I think, rightly so. (If I were choosing whether to try to initiate a romantic relationship with Carla or with Daphne, I hope that my own sense of beauty would carry more weight than conventional standards.)

Could Hume's theory of taste allow us to make sense of the idea that some desires are erroneous? Certainly, a meaningful concept of a *desirable* object can be defined in terms of properties of that object, as viewed from a conventionally fixed point. Desirability might be defined in relation to the normal desires of people in general (as when an old-fashioned property agent describes a house as a 'desirable residence'), but if we are looking for connections with behavioural economics, it is more useful to define it in relation to the normal desires of a specific individual. Think how the concept of preference is used in ordinary language. Suppose that, in general conversation, I am asked whether I prefer tea or

coffee as a breakfast drink. On a natural interpretation, I am being asked which I *normally* desire more when I am having breakfast. In other words: What desires are induced in me by the normal cues of breakfast situations? Suppose the correct answer to that question is that I normally desire tea rather than coffee: in the sense of the question, I prefer tea. Nevertheless, there are occasional breakfasts at which I feel a desire for coffee rather than tea, and so choose to drink coffee. That desire might be prompted by what Hume calls momentary appearances – perhaps the smell of someone else's coffee happening to activate some happy memory with coffee associations. Have I made an error? I think not. While drinking the coffee, I can still say that my normal desire is for tea, and that in stating this as my preference, I am correcting for momentary context-dependent cues. But at that moment, my actual desire is for coffee, and that desire is not in need of correction. The existence of context-independent standards of correctness in judgements about desirability does not imply that there are corresponding standards of correctness in desires.

3.5 Interest

When Hume is developing his theory of mind – that is, in Book I ('Of the understanding') and Book II ('Of the passions') – he occasionally uses the concept of a person's *interest*. As this concept is central to the theory of justice that he proposes in the third and final Book ('Of morals'), it is worth asking whether error might be defined as acting contrary to one's interest. Pursuing this question, however, one discovers that 'interest' – unlike 'impression', 'idea' or 'passion' – is not a technical term in Hume's theory of mind, and is never defined. Pursuit of interest is not one of the many passions that get specific treatment in Book II. Hume uses 'interest' only in asides to his theory, presumably intending it to be read in the ordinary-language sense that was current in the mid-eighteenth century.

Here are some typical examples. Describing 'vulgar lying', Hume contrasts interest with vanity: 'men without any interest, and merely out of vanity, heap up a number of extraordinary events, which are either the fictions of their brain, or if true, have at least no connexion with themselves' (p. 301). Discussing gambling, he contrasts interest with entertainment: 'the pleasure of gaming arises not from interest alone; since many leave a sure gain for this entertainment' (p. 452). Discussing the desire for knowledge, he contrasts interest with idle curiosity: 'Some people have an insatiable desire of knowing the actions and circumstances of their neighbours, tho' their interest be no way concern'd in them' (453). Discussing a case in which we admire the beauty of someone else's house, he contrasts interest with the appreciation of beauty: 'But after what manner does it give pleasure? 'Tis

certain our own interest is not in the least concern'd' (p. 364). Hume also hints at contrasts between interest and duty (p. 2), interest and honour (p. 382), interest and love (p. 356), and interest and the desire to punish (p. 418).

Notice how benevolence, malevolence, duty, honour and love are all contrasted with interest. The implication is that 'interest' is interpreted as *self*-interest. Notice also that, in these examples, *taking* pleasure or *satisfying* desire is never treated as promoting a person's interest. This may seem surprising, given that the accumulation of personal wealth (surely an object of one's interest) is *a means to* pleasure, as Hume recognises: 'The very essence of riches consists in the power of procuring the pleasures and conveniences of life' (p. 315). I suggest that Hume's concept of a person's present interest is best understood in terms of her power to satisfy the self-interested desires she can expect to have in the future. It would be reasonable to assume that, whatever other desires they may have, most people have calm – and sometimes even violent – desires to further their own interests. But that is not to say that there is an unambiguous, one-dimensional measure of interest. It seems more credible to assume that, just as people have multiple calm desires, so they have multiple interests that they may weight in different non-erroneous ways, depending on the psychological cues to which they are exposed. In any case, Hume's concept of interest seems to exclude too much of what people do in fact desire to serve as a general criterion of correctness of preferences.

4. Justice and interest

A large part of Hume's Book III is taken up by an analysis of three fundamental rules of *justice*. These are rules for the stability of possession of goods, the transference of possession by consent, and the performance of promises. According to Hume, society is absolutely necessary for human well-being, and for society to function, it is just as necessary that these rules are observed (p. 526). Since Book III is about morals, Hume's task is to provide a psychological explanation of why these rules are perceived as moral obligations – of '*Why we annex the idea of virtue to justice, and of vice to injustice*' (p. 498).

Hume (pp. 477–484) argues that the explanation cannot be that human beings have a natural desire – a desire implanted by Nature or, in modern language, one that is biologically hard-wired – to observe the rules of justice, or to approve of other people's observance of them. He points to the differences between justice on the one hand and benevolence or 'humanity' on the other, conceived as virtues. An act of humanity – for example, a rich man giving aid to someone in need – can be immediately perceived as humane, without

characterising it as the observance of any general rule. It is reasonable to explain such acts as prompted by natural human sympathy, activated by context-specific cues of contiguity: '[T]here is no human, and indeed no sensible, creature, whose happiness or misery does not, in some measure, affect us, when brought near to us, and represented in lively colours' (p. 481). In contrast, the virtue of an act of justice – for example, repaying a debt – derives from the act's relationship to a system of socially contingent rules. Hume concludes that the existence of those rules must be conceptually prior to the perception that they impose moral obligations. (That is why he can talk about *annexing* the idea of virtue to justice.) Accordingly, he proposes a non-moral theory that can explain how rules of justice can come into existence and why people comply with them. My concern is with this theory.

Having presented this theory, Hume declares: 'The *natural* obligation to justice, *viz.* interest, has been fully explained' (p. 498). And: '*Thus self-interest is the original motive to the* establishment *of justice*' (p. 499). As these statements imply, the theory is constructed in a conceptual framework in which 'interest' is a fundamental component. It is based on a model of human psychology that is much simpler than the theory of mind developed in Books I and II, but which is sufficient for Hume's purposes.

> Here is how Hume describes the essential assumptions of his model:
>
> I have already observ'd, that justice takes its rise from human conventions; and that these are intended as a remedy to some inconveniences, which proceed from the concurrence of certain *qualities* of the human mind with the *situation* of external objects. The qualities of the mind are *selfishness* and *limited generosity*: And the situation of external objects is their *easy change*, join'd to their *scarcity* in comparison of the wants and desires of men. (p. 494)

Hume is assuming that there are goods ('external objects') that can be possessed by individuals and that can easily move from one person's possession to another, either by joint actions in which one person gives and the other takes, or by unilateral actions of taking. Individuals have desires to possess these goods. Because goods are scarce, these desires cannot all be satisfied. The assumption that there is limited generosity, as opposed to pure self-interest, is not strictly necessary for his model. Its function, I think, is to head off the criticism that he is imagining a Hobbesian state of nature: he acknowledges that most people feel generosity towards their family, friends and acquaintances, but argues that this feeling is too relationship-specific to be the foundation of justice (pp. 486–487). Two further assumptions are implicit in Hume's analysis of this model: engaging in and defending against unilateral taking are costly (which accounts for the importance of the rule of stability of

possession), and exchanges of goods between people are often desired by both parties (which accounts for the importance of the rules about transfers of possession and performance of promises).

Hume argues that, in a sufficiently small society with the properties of his model (or, as Hume says, 'on the first formation of society' [p. 499]), the three rules of justice will emerge as *conventions*. For Hume, a convention is a general rule that governs interactions between the members of some group of people, with the property that general adherence to the rule works to everyone's benefit, and that it is in each individual's interest to observe the rule, conditional on the expectation that the others will observe it too.

Here is Hume's famous description of the convention of stability of possession:

> This convention … is only a general sense of common interest; which sense all the members of the society express to one another, and which induces them to regulate their conduct by certain rules. I observe, that it will be for my interest to leave another in the possession of his goods, provided he will act in the same manner with regard to me. He is sensible of a like interest in the regulation of his conduct. When this common sense of interest is mutually express'd, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be call'd a convention or agreement betwixt us, tho' without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are perform'd upon the supposition, that something is to be perform'd on the other part. […] Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. (p. 490)

And here is Hume on the convention of performance of promises:

> Hence I learn to do a service to another, without bearing him any real kindness; because I forsee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv'd him, and he is in possession of the advantage arising from my action, he is induc'd to perform his part, as foreseeing the consequences of his refusal. (p. 521)

Hume's main theoretical conclusion is that, in a small society in which the rules of justice are generally observed, it is each individual's interest to observe those rules. He does not tell us exactly how this conclusion is derived. Part of the argument is summarised in the claim that 'the whole plan or scheme [of justice] is highly conducive, or indeed absolutely requisite, both to the support of society, and the well-being of every individual' (p. 497). In

other words, everyone benefits from everyone's observance of the rules of justice. But Hume also needs to show that this mutually beneficial scheme is an equilibrium – that it is in no one's interest to break the rules unilaterally. Hume claims that this is the case, but equivocates about why. Sometimes he seems to use the *instability argument* that any breach of a rule of justice, even in a relatively large society, is liable to endanger the whole scheme; this fact is merely more obvious in a small society (p. 499). At other times, as in the passage 'Hence I learn …', quoted above, he uses the *reputation argument* that, were any individual to violate a rule of justice in any specific instance, that individual would incur a serious risk that other people would not observe the rule in future interactions *with her*, thus excluding her from benefits that are absolutely requisite to *her* well-being.

This result is not quite enough for Hume's purposes. He still needs to show that, starting from a state in which the rules of justice are *not* observed, the society of his model will gravitate to a state in which they *are*. His remarks about rules 'arising gradually and by a slow progression' suggest that he has in mind some process of social evolution.[9]

For my present purposes, it is sufficient to consider what Hume *claims* to have shown. He has presented a simple but credible model of self-interested behaviour in a small society. In this society, the rules of justice emerge from repeated interactions between individuals. Hume interprets this result as support for his claim that the natural obligation to justice is interest. It is implicit in this claim that 'interest' is defined in terms of the kinds of desires that Hume needs to attribute to the actors in his model – essentially, their desires to possess goods.

Notice that, up to this point, Hume's analysis of justice has been almost entirely independent of his theory of mind.[10] This is not an inconsistency in the overall argument of the *Treatise*. Hume has been able to use a simple and abstract concept of individual interest because he has needed to apply it only to a very general object – a system of rules that

---

[9] On the face of it, this hypothesis about how the rules of justice emerge is consistent with the reputation argument but not with the instability argument. If the justice equilibrium is highly unstable, it is hard to see how it could be the end point of an evolutionary process. In Sugden (1986) I propose a reconstruction of Hume's model of the evolution of the rules of justice, using theoretical methods adapted from evolutionary biology.

[10] As part of his analysis of justice, Hume (pp. 501–513) addresses the question of how property rights are in fact determined. Here he does use his theory of mind, but not in relation to desire. The essential idea is that, because property is a relation between a person and an object, there is a natural tendency to associate property rights with other psychologically salient person–object relations, such as contiguity and causation. In this analysis, Hume anticipates Schelling's (1960) theory of focal points. On this, see Sugden (1986: 91–97).

governs a whole class of interactions across a society. Further, the rules of justice are themselves very general. They provide individuals with opportunities that can be used in many different ways, for many different purposes, at many different times. For a person to recognise that, over an expanse of future time, she is likely to be able to benefit by making credible promises or by engaging with other people in voluntary exchanges, she does not need to predict the particular promises or exchanges she might want to make. Nor does she need to believe that her future actions will be the product of integrated preferences: she may recognise that her preferences will be influenced by context-specific psychological cues. It is sufficient that she wants it to be the case that, in whatever situations arise in the future, she is able to get what she then wants.[11]

5. Time-preference reversal and the concept of error

Having arrived at his result about the natural obligation to justice, Hume expresses some embarrassment about it. If that obligation is as strong as he has claimed, why do societies need police and judicial systems to enforce the rules of justice? Since the section of the *Treatise* in which this question is posed is called 'Of the origin of government', Hume clearly sees it as equivalent to: Why do societies need civil government? As he puts it:

> Since, therefore, men are so sincerely attach'd to their interest, and their interest is so much concern'd in the observance of justice, and this interest is so certain and avow'd; it may be ask'd, how any disorder can ever arise in society, and what principle there is in human nature so powerful as to overcome so strong a passion, or so violent as to obscure so clear a knowledge? (p. 534)

It seems that at least one of the assumptions from which the result about natural obligation was derived must have been too strong. It is at this point that Hume draws on his theory of mind in an attempt to identify the principle in human nature that makes government necessary. Since I will be subjecting Hume's argument to close scrutiny, I need to state it in his exact words.

I begin with the most fundamental part of the argument, in which Hume makes a connection with his theory of mind. This part of the argument is not specific to justice; it is about what would now be called *time preference*. It is particularly interesting to the modern

[11] This concept of opportunity, and the claim that individuals can recognise the value of opportunities so defined, is developed and defended in Sugden (2018).

reader because it seems to invoke concepts of correctness and error in preferences, similar to those used by many behavioural economists:

> When we consider any objects at a distance, all their minute distinctions vanish, and we always give the preference to whatever is in itself preferable, without considering its situation and circumstances. This gives rise to what in an improper sense we call *reason*, which is a principle, that is often contradictory to those propensities that display themselves upon the approach of the object. In reflecting on any action, which I am to perform a twelve-month hence, I always resolve to prefer the greater good, whether at that time it will be more contiguous or remote; nor does any difference in that particular make a difference in my present intentions and resolutions. My distance from the final determination makes all those minute differences vanish, nor am I affected by any thing, but the general and more discernable qualities of good and evil. But on my nearer approach, those circumstances, which I at first over-look'd, begin to appear, and have an influence on my conduct and affections. A new inclination to the present good springs up, and makes it difficult for me to adhere inflexibly to my first purpose and resolution. This natural infirmity I may very much regret, and I may endeavour, by all possible means, to free my self from it. I may have recourse to study and reflexion within myself; to the advice of friends; to frequent meditation, and repeated resolution: And having experienc'd how ineffectual all these are, I may embrace with pleasure any other expedient, by which I may impose a restraint upon myself, and guard against this weakness. (pp. 536–537)

The final sentences about freeing oneself from a mental weakness by imposing restraints on oneself comes close to anticipating Thaler and Sunstein's (2008: 5) argument that imperfect self-control leads individuals to make 'pretty bad decisions', and that the findings of behavioural economics point to ways in which people can be steered towards choices that make them 'better off, *as judged by themselves*'.

At the core of this passage is one Hume's most brilliant experiments. In both its design and its results, it anticipates experimental research done by behavioural economists more than two hundred years later.[12] Suppose I think about prospects of monetary gain, now and in the future. Holding constant the date at which gains will appear, I desire larger gains more than I desire smaller ones. Suppose I am asked what gain of money tomorrow is equivalent to £100 today. I might say: £102. If, in 365 days' time, I am *then* asked what gain in one day's time is equivalent to £100 then, I will give a similar answer. But if I am asked *now* what gain in 366 days' time is equivalent to £100 in 365 days' time, I will probably give

[12] One of the earliest of these experiments was reported by Thaler (1981). Frederick et al. (2002) review the huge amount of work done up to 2002.

an answer much closer to £100. The implication is that, as the date at which gains will be received approaches, there is a systematic tendency for *time-preference reversal*: smaller-sooner options are increasingly likely to be preferred to larger-later ones.

This is a significant fact of human psychology for which Hume's theory of mind provides a convincing explanation – namely, that the vivacity of a desire for an object depends on contiguity, and perceptions of contiguity are subject to diminishing sensitivity. There is a hint in Hume's account that this effect might be analogous with the effect of physical distance on our visual perception of the relative size of objects, which would imply hyperbolic rather than exponential discounting.[13] But Hume's discussion of this effect introduces additional ideas that seem incompatible with his theory of mind. The concepts of 'whatever is in itself most preferable' and 'the greater good' are unexplained. Preference, for Hume, is a matter of desire, and desire is an original fact and reality, complete in itself; it does not refer to any properties of the object that is desired. All we can say is that the vivacity of desire depends on contextual cues. One might reasonably assume that at any given time, most people would prefer a larger monetary gain accruing at some given (current or future) time to a smaller gain accruing *at the same time*, and that in that sense, *money* is desirable in itself; but that tells us nothing about how gains of money should be discounted over time. The same would apply to any other one-dimensional measure of gain, even (on the assumption that such a measure is possible, which Hume doubts) to a one-dimensional measure of pleasure. The fact that pleasure is desirable in itself tells us nothing about how gains of pleasure should be discounted over time.

Here Hume seems to be slipping from an analysis of passion to an analysis of taste. Recall that, according to Hume's account, judgements of taste are governed by conventions that fix the viewpoints from which those judgements are to be made. For at least some types of judgement, it seems right to say that it is conventional to use distant viewpoints. (Hume's example of the beauty of a face suggests that this is not a universal convention.) But recall Hume's remark about not giving a shilling towards the rebuilding of someone else's inconvenient house: judgements of taste do not control our passions unless they touch the heart. As one gets closer to a potential object of desire, passion tends to displace taste. But if the passion differs from the taste, that is not an error.

---

[13] The visual angle of a line of fixed length is proportional to the inverse of its distance from the eye, i.e. distance is discounted hyperbolically. The apparent relative size of larger-further and smaller-nearer objects changes (and diverges increasingly from actual relative size) as one approaches them.

What about Hume's concept of *resolve*? A resolution is a future-directed mental state that is intermediate between desire and volition. Presumably it too is complete in itself. When the time to act on a resolution arrives, the memory of having made it may, by an association of ideas and impressions, cause a desire to act on it. But, equally, it may not: since the vivacity of memory decays over time, a desire for a present good may override it.

There is nothing inconsistent in Hume's claim that, at some point in time, a person might in fact form a resolution, and try to bind himself to act on it. But if, at a later time, that person was able to escape that commitment and chose to do so, would that be an error? Recall that Hume's 'improper' sense of reasonableness allows an action to count as unreasonable if it is based on a false belief about a matter of fact or about cause and effect. If that concession is to apply to the present case, the person who fails to act on his resolution must be acting on some false belief. But when Hume applies his analysis of time preference to justice, he says:

> This is the reason why men so often act in contradiction to their known interest; and in particular why they prefer any trivial advantage, that is present, to the maintenance of order in society, which so much depends on the observance of justice. (p. 535)

The implication seems to be that the individual's resolve to choose the larger-later option is based on his belief about what it will be in his interest to do. When the time to act arrives, he does not change his beliefs; he *knowingly* chooses what his not in his interest. The individual's thought might be put into words as: 'I know that observing the rules of justice is in my long-term interest. But just now, my long-term interest doesn't feel as important to me as the present advantage. So I am choosing the present advantage.' This is not an error of belief. Could it be *akrasia* – weakness of will? In Hume's account, as I read it, the individual who chooses the present advantage acts with an inner sense of volition, just as he did when he formed his resolve. This is a *change* of will, not a *failure* of will. I cannot see how Hume's theory of mind can support any other interpretation.

Consistently with that theory, Hume might say that the person who fails to keep his resolution is showing a lack of *virtue*. Hume's theory of virtue is a branch of his analysis of taste (discussed in Section 3.4 above). A virtue is an ability or trait of character of one person that tends to induce sentiments of approval in others. Just as a well-designed house elicits the disinterested approval of bystanders, so a character trait that confers 'advantage in the conduct of life' elicits disinterested sentiments of esteem. Hume's list of such virtues

includes perseverance, constancy and resolution (pp. 610–611). But recall again the case of the ill-designed house: a lack of alignment between taste and passion is not an error.

I conclude that, contrary to what Hume seems to be saying, his analysis of time-preference reversals does not support the conclusion that individuals are switching from earlier true preferences to later erroneous ones. All we can say is that their preferences are context-dependent, and so change over time. However, it is not clear that Hume needs to invoke time-preference reversals in order to explain why individuals might choose to have the rules of justice enforced by government. Let me explain.

6. Self-imposed constraints and the origin of government

Hume's account of the origin of government treats government as a solution to a problem caused by the human tendency to choose present advantages rather than the pursuit of long-term interests. Here is how he describes this problem:

> The consequences of every breach of equity seem to lie very remote, and are not able to counterballance any immediate advantage, that may be reap'd from it. They are, however, never the less real for being remote; and as all men are, in some degree, subject to the same weakness, it necessarily happens, that the violations of equity must become very frequent in society, and the commerce of men, by that means, be render'd very dangerous and uncertain. You have the same propension, that I have, in favour of what is contiguous above what is remote. You are, therefore, naturally carried to commit acts of injustice as well as me. Your example both pushes me forward in this way by imitation, and also affords me a new reason for any breach of equity, by shewing me, that I should be the cully [i.e. dupe] of my integrity, if I alone shou'd impose on myself a severe restraint amidst the licentiousness of others. (p. 535)

The solution to this problem, according to Hume, is the establishment of a civil government, responsible for the execution of justice. Each individual accepts this constraint on his own action, conditional on every other individual's acceptance of the same constraint (pp. 537–539).

Recall, however, that in Hume's analysis of the general problem of time-preference reversals, there is no interlocking of self-imposed constraints. Each individual recognises *his own* tendency to prefer present advantage over long-term interest, and construes this as a private weakness that, in principle, could be corrected by a *unilateral* act of self-restraint. It is puzzling that Hume uses this case to model the role of government in enforcing the rules of justice.

For Hume's argument about the origin of government to work, it is sufficient to assume that, in the absence of government, individuals are sometimes tempted to break the rules of justice; it does not matter what the temptation is. These occasions of temptation might be ones in which breaking a rule really is in an individual's interest (perhaps because the probability of being discovered breaking the rule is very low); or they might be ones in which the individual *believes* that breaking a rule would be in his interest, even though that belief is in fact false; or (as in the cases Hume considers) they might be ones in which the individual's desire for present advantage leads him to break a rule, knowing that this act is contrary to his interest. Whichever of these cases applies, it seems clear that it is in each individual's interest that *other individuals* do not act on *their* temptations to break the rules of justice. As Hume puts it: '[W]e never fail to observe the prejudice we receive, either mediately or immediately, from the injustice of others; as not being in that case either blinded by passion, or byass'd by any contrary tempation' (p. 499). It is therefore in each individual's interest that other individuals are subject to the government's enforcement of justice. As Hume recognises when he talks about being the cully of one's integrity, it may not be in an individual's interest to subject himself to such constraints when others who are not subject to them might be tempted to behave unjustly. The implication is that, for Hume, the institution of government is a convention of *mutual* restraint, not a collection of unilaterally self-imposed constraints. In Humean language: I observe that it will be for my interest to allow the government to compel me to act justly towards other people, provided they will allow the government to compel then to act justly towards me.

On this issue, at least, I think the older Hume is a better guide. Discussing justice in the *Enquiries*, Hume says:

> [A] man, taking things in a certain light, may often seem to be a loser by his integrity. And though it is allowed that, without a regard to property, no society could subsist; yet according to the imperfect way in which human affairs are conducted, a sensible knave, in particular instances, may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union and confederacy. That *honesty is the best policy*, may be a good general rule, but is liable to many exceptions; and he, it may perhaps be thought, conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions. (1748-51, p. 283)

The Hume of the *Enquiries* never claims that honesty *is* always the best policy. However, he points to the risks that the knave runs in trying to identify the exceptional cases, telling us that an honest man has:

> the frequent satisfaction of seeing knaves, with all their pretended cunning and abilities, betrayed by their own maxims; and while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give into the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all future trust and confidence with mankind. (1748-51, p. 283)

One might still perhaps argue that a sufficiently sensible knave would factor all this into his calculations and choose unilateral self-constraint. But if one wants to show the knave that government works to his advantage, it is surely simpler and more direct to point to the advantages that he derives from the constraints that government imposes on others, and to explain to him that he cannot have those advantages unless he too is constrained. That honest men derive satisfaction from the thwarting of knavery is part of the explanation of why even they would be reluctant to subject themselves to the enforcement of justice unless the knaves subjected themselves too.

7. Conclusion

If behavioural economists were to look for a patron philosopher, Hume would be the obvious candidate. As I have tried to show, he was a pioneer of experimental psychology. As a philosopher, he shows us how we can think clearly about our own thinking while recognising that our processes of thought are ultimately governed by principles of psychology. If Hume is right, thinking in this way requires us to give up many pre-scientific beliefs about human rationality. Even if, on rare occasions, Hume seems to suggest otherwise, I think we must accept that one of the beliefs that has to be given up is that there are standards of correctness for preferences.

References

Berg, N., Gigerenzer, G., 2010. As-if behavioral economics: neoclassical economics in disguise? *History of Economic Ideas* 18, 133–166.

Bernheim, D., 2016. The good, the bad, and the ugly: a unified approach to behavioural welfare economics. *Journal of Benefit-Cost Analysis* 7, 12–68.

Bernheim, D., Rangel, A., 2009. Beyond revealed preference: choice- theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics* 124, 51–104.

Bershears, J., Choi, J., Laibson, D., Madrian, B., 2008. How are preferences revealed? *Journal of Public Economics* 92, 1787–1794.

Binmore, K., 1994. *Game Theory and the Social Contract, Volume 1: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, K., 1998. *Game Theory and the Social Contract, Volume 2: Just Playing*. Cambridge, MA: MIT Press.

Bleichrodt, H., Pinto-Prades, J.-L., Wakker, P., 2001. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science* 47, 1498–1514.

Bolton, G., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity and competition. *American Economic Review* 90, 166–193.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., Rabin, M., 2003. Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review* 151, 1211–1254.

Demeter, T., 2012. Hume's experimental method. *British Journal for the History of Philosophy* 20, 577–599.

Frederick, S., Loewenstein, G., O'Donoghue, T., 2002. Time discounting and time preference: a critical review. *Journal of Economic Literature* 40, 351-401.

Garrett, D., 2015. *Hume*. Abingdon: Routledge.

Hume, D., 1739-40/1978. *A Treatise of Human Nature*. Oxford: Oxford University Press.

Hume, D., 1748-51/1975. *Enquiries concerning Human Understanding and concerning the Principles of Morals*. Oxford: Oxford University Press.

Infante, G., Lecouteux, G., Sugden, R., 2016. Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology* 23, 1–25.

Kahneman, D., 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291.

Kőszegi, B., Rabin, M., 2007. Mistakes in choice-based welfare analysis. *American Economic Review* 97, 477–481.

Le Grand, J,. New, B., 2015. *Government Paternalism: Nanny State or Helpful Friend?* Princeton, N.J.: Princeton University Press

Lewis, D., 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Manzini, P., Mariotti, M., 2012. Categorize then choose: Boundedly rational choice and welfare. *Journal of the European Economic Association* 10, 939–1213.

Rabin, M., 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83, 1281–1302.

Robinson, T., Berridge, K., 1993. The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Research Reviews* 18, 247–291.

Salant, Y., Rubinstein, A., 2008. (*A*, *f*): choice with frames. *Review of Economic Studies* 75, 1287–1296.

Savage, L., 1954. *The Foundations of Statistics*. New York: Wiley.

Schelling, T., 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sugden, R., 1986. *The Economics of Rights, Cooperation and Welfare*. First edition. Oxford: Blackwell. (Second edition 2004. Basingstoke: Palgrave Macmillan.)

Sugden, R., 2006. Hume's non-instrumental and non-propositional decision theory. *Economics and Philosophy* 22, 365–391.

Sugden, Robert, 2018. *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.

Sunstein, C., Thaler, R., 2003. Libertarian paternalism is not an oxymoron. *University of Chicago Law Review* 70, 1159–1202.

Thaler, R., 1981. Some empirical evidence on dynamic inconsistency. *Economics Letters* 8, 201–207

Thaler, R., Sunstein, C., 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.