Behavioural Public Policy (2024), **8**, 197–211
doi:10.1017/bpp.2022.2

CAMBRIDGE
UNIVERSITY PRESS

ANNUAL LSE BEHAVIOURAL PUBLIC POLICY LECTURE

What is the point of behavioural public policy? A contractarian approach

Nick Chater

Behavioural Science Group, Warwick Business School, Coventry, UK
E-mail: nick.chater@wbs.ac.uk

(Received 18 January 2022; accepted 18 January 2022; first published online 25 March 2022)

Abstract

Many proponents of behavioural public policy work within a broadly consequentialist framework. From this perspective, the ultimate aim of public policy is to maximise utility, happiness, welfare, the satisfaction of preferences, or similar; and the *behavioural* aspect of public policy aims to harness a knowledge of human psychology to make this maximisation more effective. In particular, behavioural insights may be crucial to help policymakers 'save us from ourselves' by helping citizens avoid falling into non-rational choices, for example, through framing effects, failures of will-power, and so on. But an alternative reading of the psychological literature is that human thoughts and actions are not biased from a rational standard, but are simply systematically inconsistent. If so, then utility and similar notions are not well defined either for individuals or as an objective of public policy. I argue that a different, contractarian viewpoint is required: that the determination of public policy is continuous with the formation of agreements we make with each other at all scales, from momentary social interactions, to linguistic and social conventions, to collective decisions by groups and organisations. Behavioural factors do not over-ride, but can (among many other factors) inform, our collective decision-making process. The point of behavioural insights in public policy is primarily to inform and enrich public debate when deciding the rules by which we should like to live.

Keywords: contractarianism; consequentialism; utilitarianism; public policy; behavioural economics

Introduction

Why does the nature of human behaviour matter for public policy? The standard answer is *descriptive*. A more accurate description of human behaviour will help policy-makers achieve their goals. But there is another answer which is less discussed but at least as important: that understanding behaviour is crucial from a *normative* point of view: to help establish what it is that policy-makers should be trying to achieve.

The idea that behavioural factors might influence a normative account may seem inherently misguided. Doesn't this require inferring an 'ought' (what our policy goals should be) from an ‘is’ (how human behaviour works), which has been viewed as fallacious since Hume (1739/1894)? Not at all, as we shall see. The relationship between human behaviour and normative considerations is more subtle than merely inferring what ought to be from how things are. But it is of central importance to the formulation of public policy.

Let us start by considering two contrasting, and influential, normative frameworks for establishing the normative basis of public policy decisions. The first viewpoint is that public policies should be judged by their consequences. This is the tradition of utilitarianism of Bentham, Mill, and Edgeworth, in which the goodness of consequences is measured in some presumed units of utility, welfare, value, pleasure or pain, or perhaps in terms of self-reported life-satisfaction (Sen & Williams, 1982). Consequentialism, for these purposes, also includes the tradition of welfare economics, in which utility is ‘revealed’ by choices (in the modern economic tradition, utility is no more than a convenient summary of preferences, such that the option or object with the higher utility is, by definition, that which is preferred).

However we conceive of utility, there remains the challenging question of comparing utility between individuals (Hammond, 1991) – which seems crucial if we are to implement some variant of Bentham’s universal yardstick of maximising ‘the greatest happiness of the greatest number.’ Without some standard of comparison (ideally on a continuous, cardinal, numerical scale), it is not clear how to trade off the happiness of one person against that of another – and there are few public policy decisions that do not involve such trade-offs. In the rare cases in which no trade-offs are required, it is often reasonable to assume that one option is ‘better’ than another at the level of society, if it is preferred by at least one person, and not dis-preferred by anyone (so that perplexing questions of the interpersonal comparison of utility can be avoided). And this approach can be extended, moreover, using somewhat controversial principles such as the Kaldor–Hicks criterion, by which an option is preferred if it would be preferred in the above sense, *if* money were redistributed appropriately (whether or not such a redistribution is undertaken). This approach is frequently applied in cost–benefit analysis in policy formulation, such as in the principles built into the UK government’s Green Book (Hurst, 2019). Finally, the consequentialist framework is in play at a national level when policies are evaluated by their effect on aggregate quantities, such as GDP, or more radically, some measure of Gross National Happiness (e.g., Frijters *et al.*, 2020).

The second normative conception of public policy focuses instead not on consequences but on agreement, whether real or hypothetical. This contractarian tradition is rooted in political philosophy, can be traced from Hobbes and Rousseau through to Rawls, and has been extended to ethics (Gauthier, 1987; Scanlon, 2000). The contractarian tradition also meshes naturally with the practical processes of politics and law. Here, the policy-maker’s objectives, and constraints on how those objectives can be pursued, are set by legislation and regulations typically designed by the executive branch of government and approved by parliament. The normative status of a public policy is not justified primarily by a consideration of its consequences, but on its *legitimacy*: a policy should be followed because it has been agreed by the legitimate authority or process (e.g., a referendum, a vote in parliament, the judgement of a relevant court, and so on). And what counts as a legitimate authority typically rests on prior agreements by other legitimate authorities (e.g., a regulatory body may have

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

been established by parliament). There seems a danger of infinite regress, of course. While the consequentialist approach aims to end any regress by appeal to some bedrock of 'basic' utility judgements (whether coming from hedonic states, well-being judgements, or individual preferences), the contractarian approach typically seeks grounding in a hypothetical 'grand bargain,' in which the rules and objectives of society are imagined to be agreed by its citizens as a whole (perhaps behind a 'veil of ignorance' to block the influence of personal interests, Rawls, 1971). We will see below that a myriad hypothetical 'local bargains,' rather than a single grand bargain, may also be an appropriate starting point.

It is uncontroversial that descriptive behavioural facts can play an important role in both these normative frameworks. For example, the earliest form of utilitarianism, Bentham's proposal that we maximise pleasures and minimise pains immediately raises the important behavioural (or more broadly, psychological) question of what leads to pleasures or pains. Equally, a preference-based consequentialist relies on descriptive facts about what people prefer. Likewise, contractarian approaches depend on descriptive facts about what people will actually agree. These descriptive facts will all, of course, be strongly influenced by psychological factors. In both cases, behavioural factors need to be 'fed into' the normative account. Hence, there is clearly a role for behavioural science, psychology, and perhaps even neuroscience, in helping to measure pleasures, pains, preferences, or agreements in a reliable way.

In this paper, I argue for the particular importance of two proposed features of human behaviour that impact on normative accounts, whether consequentialist or contractarian, in a much more fundamental way. The first focuses on the constructive, improvised nature of thought and behaviour – which implies that there are, in general, no well-defined answers to questions about 'what people really want' or 'what they would agree under ideal conditions' or similar. The second stresses that the process of constructing norms is fundamentally social – so that an individualistic foundation for public policy is not viable. Both points have significant practical implications for the normative basis of behavioural public policy.

The constructive, improvised nature of thought

A consequentialist approach to public policy typically aims to give people what they want. But this very normative statement makes a crucial behavioural presupposition: that there is a well-defined answer to the question of what it is that people want. Yet we each have daily personal experience of being unsure what we do want: our frequent struggles to 'make up' our minds seem to suggest that we often don't know our own preferences. And there is something troubling about a model of policy formation which requires the policy-maker to know our own minds better than we do ourselves. Thus, to take a recent and high-profile example, public attitudes to mask-wearing and vaccination were, both before and during the pandemic, remarkably unstable and ill-defined (Mills *et al.*, 2020). But this instability is the norm rather than the exception, as we shall see.

Our intuitions are not merely partial, but also typically inconsistent. Indeed, the entire fields of judgement and decision-making research and behavioural economics exemplify this point (e.g., Kahneman & Tversky, 2000; Kahneman, 2011). Often, the

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

researcher presents the ‘same’ problem to the unsuspecting participant in two different ways (e.g., framed in terms of losses vs. gains; by grouping options together in different, though logically equivalent, ways; adding ‘irrelevant’ options to the choice set; unpacking events into a finer-grained, but equivalent, description, and so on). The participant duly obliges by giving different answers to these versions of the ‘same’ question, which is taken to reveal that a ‘behavioural bias’ is at work.

Yet, in reality, what is revealed is better viewed as systematic inconsistency. If people ‘really’ knew what they believed or wanted, they would surely not be victim to such variations (or only rarely, due to the occasional ‘trembling hand’ error, where they inadvertently select an option they didn’t intend to choose). And their inconsistencies would surely evaporate as soon as the equivalence of the different problem formulations was pointed out. Often, though, the opposite is the case: indeed, puzzles such as Allais’ paradox and Ellsberg’s paradox (Allais, 1953; Ellsberg, 1961) are puzzling precisely because many people (including many normatively oriented researchers) wish to maintain their endorsement of inconsistent intuitions about apparently equivalent problems, even after sophisticated reflection. For that matter, if people knew what they really wanted, they should at least be consistent when repeatedly asked the *very same question*. Yet, in typical choice experiments, repetitions (after judicious intervals, but in the same experimental session) have long been known to produce remarkably high levels of inconsistent responses (e.g., Mosteller & Nogee, 1951).

To see what is going on here, it is useful to focus on a specific, but revealing, example, stemming from Eldar Shafir’s work on ‘reason-based’ choice (Shafir, 1993; Shafir *et al.*, 1993). Shafir presented people with choices between a ‘middling’ option (M), with a set of satisfactory though not outstanding attributes, and a ‘conflicted’ option (C), with some good attributes and some poor attributes.

When participants are asked whether they would like to *select* M or C, they mostly choose C. When, by contrast, they are asked which option they would like to *reject*, they also mostly choose C. This is a framing effect of a particularly stark kind. But more important for the argument here is Shafir’s explanation for the phenomenon. He suggests that, in order to make choices, people aim to formulate justifications for their choices (both for themselves, but potentially also to others, e.g., Mercier & Sperber, 2011). When asked to select between C and M, people are primed to search for a positive attribute of one of the options, to justify why that option should be chosen. The conflicted option, C, has the most extreme positive features, which can be readily assimilated into such a justification – and hence C is selected. By contrast, when asked to reject an option, people are primed to find a negative feature of one of the options, which can justify its elimination: and again, option C, by virtue of its mix of good and bad attributes, most readily allows such a justification. Thus, C is rejected. The broader point is that people are not reading off a pre-specified preference in order to make these choices. Instead, they are improvising a justification ‘on the fly,’ and depending on the nature of the question, different improvised justifications will most readily come to mind.

To test this viewpoint under carefully controlled conditions, Konstantinos Tsetsos, Marius Usher, and I developed a simple experimental paradigm which we call ‘value psychophysics’ (Tsetsos *et al.*, 2012). People were presented with two or more streams

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

of numbers on a computer screen, representing samples from distinct probability distributions. Based on these samples, they then chose one of the options, from which a new sample is drawn, which determined their reward. The critical comparisons concerned distributions of numbers with the same mean, one with a small and one with a large standard deviation, corresponding to the middling and conflicted options mentioned above.

When asked to *select* one of the options, people chose the distribution with the large standard deviation. This would be expected if they focus on finding justifications for choosing this option, as would be expected according to Shafir's account. Observing or recalling a specific high number, sampled from a distribution, provides such a justification.[1]

Just as Shafir found, when people ask to *reject* an option, they most commonly reject the distribution with wide standard deviation – the very same option that was typically selected – presumably because they are now looking for low numbers to justify rejection, and the distribution with the large standard deviation also has the lower numbers. In essence, the task of selecting one of two options leads to attention being directed to high numbers; the task of rejecting one of two options leads to attention being directed to low numbers; and since the distribution with a large standard deviation has more high and low numbers, it is both accepted and rejected. This explanation was confirmed by a series of experiments that either manipulated, or measured, the amount of attention paid to different numbers in these sequences (Kunar *et al.*, 2017).

This result provides a particularly clear illustration of a general principle (reviewed at length in Chater, 2018; see also Payne *et al.*, 1993; Slovic, 1995; Gazzaniga, 2000; Hall *et al.*, 2013): that whether we are making a choice, formulating a plan, making a judgement, or articulating our beliefs, our minds are continually scrambling to put together a justification for our response from whatever information is momentarily to hand, rather than consulting a fixed 'data-base' capturing our entire stock of beliefs and preferences. Different improvisations, on different occasions, will often lead to patterns of responses that are inconsistent according to conventional rational theories. Before the process of improvisation has begun, there is no meaningful answer to the question of what we prefer or believe, any more than there is a well-defined answer to the question of what a novelist, composer, or painter will next create, before they have created it. Choices and beliefs are, from this point of view, the *endpoint* of creative cognitive activity, rather than its starting point.

This viewpoint of the mind as continually improvising answers to whatever question it is faced with, using whatever information happens to be available at the time, predicts inevitable inconsistencies between different improvisations, because these will draw on distinct fragments of information (e.g., attending to high vs. low numbers). Indeed, as soon as one inconsistency is fixed, another will doubtless be spotted, and so on indefinitely. The process of thought can be viewed as attempting to

[1] Note that this behaviour appears to indicate that people are risk seeking – they are selecting a distribution with higher variance. But when people are given *descriptions* of gambles, rather than samples from them, they tend to show risk-averse choices (although even this generalisation is by no means reliable, e.g., Stewart *et al.*, 2003; Stott, 2006) – and, of course, most economic analysis assumes that people are risk-averse for gains (as here), due to the presumed diminishing marginal utility of money. This result is an example of a more general inconsistency between risky choices made from experience versus description (Hertwig *et al.*, 2004).

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

improvise answers to questions on which our thoughts were previously ill-defined, and to find and fix the endless stream of inconsistencies in our thinking.

Once we realise that our thinking is inherently inconsistent, the idea of founding policy decisions on particular judgements or choices, and ignoring all others, seems difficult to defend. Consider, for example, the various possible ways in which we may reach decisions with environmental implications, such as whether to cut down woodland to build a road or railway. Suppose, for simplicity, we intend that our sole criterion is the costs and benefits in terms of the welfare of individual citizens. On the benefit side, we might ask individual citizens how much they would pay for a journey time to be reduced by, say, one minute; and we might then estimate the total number of journey-minutes saved per year. We might also ask citizens how much they would be willing to pay to block the development and maintain the wood (e.g., because of its amenity value); or how much they would be willing to pay to save a specific tree, and attempt to extrapolate how much they would pay to save many trees. Or we might focus not merely on trees but willingness to pay to avoid disruption to nesting sites, woodland animals, and so on. There is an almost endless number of possible ways in which one might attempt directly to elicit these welfare impacts, both positive and negative, but, in the light of our general tendency to substantial inconsistency, we should anticipate that the questions will produce inconsistent answers. Indeed, it turns out that the valuation of such 'public' goods is typically wildly dependent on the method used (e.g., Hsee & Rottenstreich, 2004).

Alternatively, we might indirectly attempt to infer people's valuations of costs and benefits by comparing with their observed propensity to pay for using toll roads or comparing house prices near wooded areas versus near large roads. This approach can also be conducted in a wide variety of ways; and it will simply add to our spectrum of possible welfare impacts (whether positive or negative), rather than yielding any definitive conclusion.

Indeed, the problem of manifold inconsistency has broader implications, even for the foundations of conventional neoclassical economics. Infante *et al.* (2016) note that even a descriptive model of economic behaviour requires the ability to reconstruct some 'purified' set of preferences from the inconsistent preferences and choices that people actually exhibit. They are, I think rightly, sceptical, that any such process of purification is possible. But if no consistent set of preferences can be distilled, then cost–benefit analysis based on welfare has no stable foundation. Indeed, Infante *et al.* (2016) suggest that the wider programme of building a 'behavioural' welfare economics which takes account of frailties of human reasoning (Sunstein & Thaler, 2003; Bernheim & Rangel, 2007; Hausman, 2012) is undermined.

A really committed utilitarian might hope that, in spite of these substantial practical difficulties, there is some answer to questions about which options (to build or not to build) have the better consequences for human welfare. For example, Edgeworth imagined the possibility of constructing a 'hedonimeter'

> '… let there be granted to the science of pleasure what is granted to the science of energy; to imagine an ideally perfect instrument, a psychophysical machine, continually registering the height of pleasure experienced by an individual …' Edgeworth (1881/1961, p. 101)

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

Edgeworth envisaged the machine to have a well-defined reading of zero, and a cardinal scale somehow calibrated to compare the momentary subjective experiences of individuals. Then, he imagined, we would have a definitive ethical objective – to maximise the total sum of utility as measured by the machine.[2]

The existence of such a device is probably not even conceptually coherent, let alone practically feasible. But even if it were, it would merely add yet another criterion for valuation and choice, which would, inevitably, be inconsistent with all the others. And there would seem no obvious reason why the maximisation of momentary positive subjective experiences should take precedence. The problem is deepened in the light of findings that people's subjective reports of utility, and indeed subjective experiences of any kind, are not typically consistent across timescales (Hausman, 2015; Oliver, 2017) and that utility judgements are not consistent when viewed in prospect, in the moment, or in retrospect (e.g., Kahneman *et al.*, 1997). And choice and valuation are frequently in conflict. This is evident in preference reversals, in which people choose options to which they assign lower values (Lichtenstein & Slovic, 1971), as well as in the fundamental psychological and neuroscientific distinction between wanting and liking (Berridge & O'Doherty, 2013), such that people often make repeated choices which do not appear to maximise their amount of 'pleasure' by any reasonable measure.

Indeed, when faced with inconsistent choices, inconsistent valuations, and yet further inconsistencies *between* preferences and choices, which should we trust as providing a solid foundation? This seems a question with no viable answer when we realise that these inconsistencies arise not from occasional 'noise' distorting a perfectly rational agent, but rather emerge because of the inherently *ad hoc*, improvised nature of thought. Thus, to return to the point with which we began, purely descriptive facts about psychology turn out to have potentially far-reaching normative implications, which may include, arguably, the wholesale demise of a consequentialist normative picture.[3]

Public policy by agreement: real bargaining and virtual bargaining

The consequentialist hopes to determine the 'right' public policy by aggregating the impacts on each individual, measured or estimated by some means (though not, perhaps, by the use of anything resembling Edgeworth's hedonimeter). This is

[2] How feasible does Edgeworth's machine look, in the light of modern neuroscience? While some advocates of neuroeconomics have suggested that neuroscience, including the application of brain imaging technology, might take us a step towards the direct measurement of utility, I suspect that the opposite may prove to be true (see, for various perspectives, Camerer *et al.*, 2005; Glimcher *et al.*, 2005; Vlaev *et al.*, 2011). There are no well-defined 'centres' for pleasure or pain, whose activity could be monitored as Edgeworth might have hoped; instead, the phenomenological experience of diverse pleasures and pains emerge from a process of appraisal involving complex patterns of activity across large parts of the brain (Melzack, 1990). Moreover, even when systematic neural responses to specific rewards or punishments can be identified neurally, these behave in ways that cannot serve as a scale of utility (having very limited precisions and dynamic ranges, and showing fast local adaptation, e.g., Tobler *et al.*, 2005).

[3] We will see below, though, that cost–benefit analysis, and consequentialist thinking, is still important, as providing a consistency criterion for what we agree. But it does not determine what our agreements should be.

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

an inherently technocratic perspective: the clever and all-knowing technocrat is tasked with establishing what will maximise the public's utility or best satisfy their preferences, whether individual citizens recognise this or not. It is also a fundamentally individualistic perspective: public policy is evaluated by aggregating its impact on each individual citizen.

A very different starting point is possible. Here, again, descriptive features of human psychology may be helpful in addressing normative questions. One of the most striking features of human society, in distinction from social organisations among other animals, is the incredible variability and flexibility of our culture and institutions, ranging from the remarkable variety of human languages (Evans & Levinson, 2009; Christiansen & Chater, 2022), to widely differing musical styles, literary traditions, social norms, regulations and laws, religious beliefs, scientific and technical knowledge, and systems of political and economic organisation (Fukuyama, 2011). It has been argued that underpinning the creation of this cumulative cultural complexity is a suite of interrelated but distinctively human cognitive mechanisms.[4] These may include the ability to create and enforce rules of behaviour (e.g., March & Olsen, 2008), a sense of obligation to follow those rules (Tomasello, 2020), the ability to attend, act and reason jointly (e.g., to ask 'what should *we* do?', rather than 'what should I do?', Colman, 2003; Sugden, 2003; Bacharach, 2006; Chater *et al.*, in press), to use language to formulate complex joint plans, hypotheses, belief systems, and much more.

Crucially, almost all the agreements relevant to social interactions are unstated. For example, participants in a conversation are guided by complex linguistic rules concerning the operation of their language (e.g., Culicover, 1999; Huddleston & Pullum, 2002), pragmatic principles of communication (e.g., Grice, 1975; Levinson, 2000), norms of social interaction and rules about roles and social status (Garfinkel, 1967), and much more.

Central to these diverse capabilities is the human ability to form agreements, both explicit and tacit. We can decide, between us, what are the appropriate social norms for our group; and by agreeing, we are automatically under some obligation to conform, and perhaps also to attempt to try to ensure that others conform. We agree not just on plans and actions, but values, assumptions, working hypotheses, and expectations. Implicit agreements concerning the meanings of words and the principles of grammar underpin our ability to formulate our thoughts, including what we have agreed, in a shared public form – ranging from dictionaries to grammar books, to maps, textbooks, financial accounts, legal contracts, rulebooks, and written constitutions. It is the ability to make agreements that make possible the construction of the public products of thought that Popper (1972) terms World 3 (in addition to the 'worlds' of natural processes and mental states).

To play a role in a society, we have to agree about a lot. Without a shared knowledge of the natural, social, and economic world, and a common language, we would be unable to coordinate our behaviour to engage in conversations, form relationships, create hunter-gather communities, sports teams, religious movements, judicial and

[4] Whether each of these elements is uniquely human is controversial, but it is generally agreed that these abilities, if present in nonhuman animals at all, are very much less well developed than in humans.

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

political systems, and so on. To disagree about anything in particular, presupposes agreement about almost everything else. To argue about the fair distribution of money (whether when paying a restaurant bill, or setting taxes), for example, presupposes a strong measure of agreement concerning the nature of money, the notion of fairness, the parties between whom the money might be divided, and a common language in which the argument might be expressed.

But while any disagreement presupposes a bedrock of common ground (so that it is incoherent to disagree about everything, simultaneously), people can and do disagree on almost any topic. Indeed, such disagreements, and the debates, rapprochements, and further divisions that they create are a driving force behind continual cultural and social change. Speakers of a language disagree about points of grammar or the meanings of particular words; academics cast doubt on the interpretation of each other's data and the coherence of each other's theories; lawyers debate the guilt of a specific defendant, or wider points of law; politicians debate legislation; and so on. From this point of view, public policy is just a particular outcome of processes of agreement and disagreement, albeit concerning agreements of especially large scope and importance.

From this point of view, then, how should we evaluate and improve public policy? And what should be the role, if any, of a specifically *behavioural* public policy? A rough starting point regarding the first question is to switch from the consequentialist's focus on policy *outcomes*, to focus on the *legitimacy* of the policy – that is, how the policy came to be imposed. The essence of the uniquely human ability to agree is that we are collectively obliged, at least other things being equal, to guide our behaviour in line with the agreement, rather than freely choosing whatever option appears to lead to the best consequences 'in the moment.' As we have noted, this ability to create and live by agreements is a bedrock of human society.

Thus, for example, in line with constitutional agreements governing a democracy, a losing party in an election is obliged to cede power, even if there might be overwhelmingly strong reasons to believe that the new government may pursue disastrous policies. Similarly, there is at least some obligation for a political party to implement its manifesto commitments, even in the face of compelling arguments that these commitments are misguided. And nations are generally presumed to be required to honour international agreements (e.g., to reduce carbon emissions), and are not free to jettison these purely on the basis of revised cost–benefit calculations.

So if we see public policy decisions as types of agreement, one natural standard against which policy decisions can be judged concerns how those policy decisions are reached. If such decisions were reached, for example, in breach of prior commitments, without following the appropriate political or legal processes, influenced by perhaps undeclared lobbying from special interest groups, or simply decided and even implemented in secret, then there is a credible *prima facie* challenge to the policy's legitimacy. But, without further constraints, this is surely far too loose a standard against which public policy should be held. It seems to imply that any policy that results from correctly following agreed decision-making procedures is by virtue of this fact beyond criticism.

But from what vantage point can a contractarian criticise an agreed policy? A consequentialist can, of course, evaluate policies directly by their results. But the

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

contractarian wants to avoid recourse to any objective standard of 'welfare' or 'utility' against which results can be evaluated – a reaction that is often based on the challenges of interpersonal comparison of utility, but as we have seen is strengthened by the psychological finding that our choices and valuations are improvised *ad hoc* and are thoroughly inconsistent (so that utility within the person is also incoherent).[5]

There is, however, an alternative move for the contractarian, which has a natural psychological basis. If a policy arises through the proper procedures, then it is, indeed, legitimately agreed – and we are therefore obliged (other things being equal) to follow its precepts. But we are nonetheless free to object that the policy should be replaced by some other policy, if we believe that some alternative policy would have been preferred, under various hypothetical circumstances. So, for example, we might propose that a policy is appropriate if the affected parties *would* have agreed to it, had they been consulted and in full possession of the relevant facts.

Suppose, for example, that we are in an economy in which there is no tax on carbon emissions. Indeed, the absence of such a tax might even be a declared plank of policy for the elected party or coalition (or perhaps even the result of a referendum in a country with Swiss-style direct democracy). The absence of a carbon tax can still reasonably be criticised by outside parties. One reason for criticism, for example, might be that the voters making a decision (whether directly or through the government they elected) were not fully aware of the impacts of carbon emissions on the climate, and the consequent implications for the ecosystem and human life. It is here that consequentialist considerations enter the contractarian account. There can and should, of course, be reasoned debate concerning both what the consequences *are* (and the relevant uncertainties); and also whether voters would have made a different decision if they had been fully aware of these consequences.

An alternative line of attack is that not all the affected parties who should have been involved in the decision were actually consulted. Suppose a western democracy votes to avoid any form of carbon tax, but the vast bulk of the harms are caused to poor countries affected acutely by extreme weather events, sea level rises, and increases to already high summer temperatures. Then a credible criticism of the lack of a carbon tax is that had all the affected parties been involved in the decision they would certainly not have agreed to it. (And, indeed, where appropriate, it is a credible criticism of international agreements that there is no mechanism for the views of these affected parties to be heard.) Here, too, standard consequentialist, cost–benefit, thinking is potentially relevant because it helps clarify when agreement might be possible, at least in principle. Suppose we consider an agreement in which a Western 'polluting' country compensates the parties negatively affected by the emissions (most likely including some of its own citizens). If the 'negative externality' is fairly modest, then such an agreement might be struck, by mutual consent. If, as is far more likely, the negative impacts are large, and hence the compensation

[5] Another possibility is to criticise the 'initial position' from which agreement was made, if the society is, for example, sufficiently unequal or otherwise unjust (e.g., Rawls, 1971). This viewpoint is parallel to the thought that laws in a sufficiently despotic or monstrous regime are not legitimate source of constraint on our actions or perhaps even do not count as laws at all [see, e.g., the debate between Hart (1958) and Fuller (1957)].

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

required is similarly large, then to obtain hypothetical agreement the carbon-emitting country would probably prefer to reduce its carbon emissions (perhaps via a carbon tax which redistributes its revenues in sum or in part to the negatively affected parties).[6] So, thinking about costs and benefits for each party to a (real of hypothetical) agreement are useful guides for establishing what either party might agree to – but an overall cost–benefit analysis aggregating the welfare impacts across the parties to the agreement is not required.

I suggest that the contractarian approach to public policy meshes with behavioural principles at its very foundation. The human mind is not a consequentialist optimiser, but an *ad hoc* solver of problems and generator of reasons, explanations, and justifications. And it is natural to see the creation of agreements as both arising from such reasoning, and as being continually under scrutiny, and even attack, from rival lines of reasoning. Human social interaction can be viewed as a matter of continual negotiation and renegotiation (Chater *et al.*, in press); decisions concerning public policy are, from the present standpoint, social interactions writ large.

In the light of this contractarian framework, the relevance of behavioural biases for public policy takes a new form. While the consequentialist sees such biases as distorting people's perceptions of what they really want, and therefore seeks to 'strip out' such biases so that people's real objectives can be revealed, the contractarian approach views the existence of behavioural biases as providing a new line of attack on, or defence of, current agreements (and hence as a providing potential arguments for reform or for maintaining the *status quo*). Thus, behavioural factors contribute to a normative account of how public policy should be set – through contributing to the debate from which, for the contractarian, our normative standards arise (see, for related discussion, Rizzo & Dold, 2020).

Consider, for example, the apparently widespread bias that we underestimate the degree to which we can successfully and happily adapt to new rules or circumstances (e.g., Brickman *et al.*, 1978; Ayton *et al.*, 2007; Wilson & Gilbert, 2013). This bias may act as a powerful conservative force when people are potentially agreeing to changes of almost any kind, from compulsory seatbelt legislation, to banning cigarettes in public places, restrictions on behaviour in response to the COVID pandemic, or taxes on carbon emissions or regulations reducing sugar and salt in food. Similarly, there appear to be powerful biases stemming from assuming others have the same beliefs or values as oneself (e.g., Camerer *et al.*, 1989), and attentional limitations that might lead to excessive focus on present concerns, the present moment, people we know personally, and so on (Kahneman, 2011).

From a contractarian point of view, knowledge of these biases can, like knowledge of any other kind, be marshalled by citizens in support of some reform to the agreements we live by, and perhaps, in favour of maintaining the *status quo*. So, for example, campaigns for more stringent legislation of almost any kind might point out to their opponents that the behavioural evidence suggests that we will all adjust

[6] This hypothetical agreement with cash transfers is, of course, close to the Kaldor–Hicks criterion in conventional welfare economics – with the exception that the Kaldor–Hicks criterion takes the acceptability to all parties of a hypothetical cash transfer as indicating that the policy can go ahead even if no cash transfer is actually made. This last step has no obvious contractarian justification.

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

to any changes much more painlessly than they expect. On the other hand, their opponents might apply similar evidence to the opposite effect: arguing that it is important not to lose freedoms, because once lost, their value is rapidly forgotten, and there will be little impetus for them to be reinstated. Similarly, many might argue that legal and political systems should actively attempt to counter the over-weighting of our own interests, and the interests of people we are close to (e.g., actively counteracting nepotism and cronyism); opponents of such reform might respond that such attempts are futile and that it is more appropriate to go with the grain of human behaviour. The point here is not that empirical evidence concerning human behaviour is entirely neutral, and can be used equally well to serve any agenda. Rather, the question of how such evidence is assimilated into the political debate is a matter for the participants in that debate, rather than something that can be resolved by a disinterested spectator, however benevolent (see Sugden, 2018).

This approach seems parallel to how we naturally treat biases in other domains. For example, suppose that a type of telescope is found to be miscalibrated or unreliable. It would seem peculiar to recommend that astronomers continue to use a telescope, report its findings, and draw their conclusions in the normal way, but that some separate groups of people should then ‘strip out’ the bias *post hoc* (and the inevitable inconsistencies that it will have led to) and then to attempt to piece together a coherent theory of the heavens from these debiased measurements. Instead, the natural strategy will be immediately to inform the astronomers, who will then presumably adapt their readings and conclusions appropriately. Similarly, in observing the ubiquitous psychological tendency to draw conclusions that are not logically valid when instructed to carry out deductive inference (e.g., Evans *et al.*, 1993), the appropriate strategy is surely not to attempt to infer people’s ‘pure’ deductive reasoning, by retrospectively correcting their reasoning errors in some way. Rather, it is to feed back this information to mathematicians and logicians themselves, so that they can be on their guard against such mistakes.[7]

So what is the point of the behavioural science in *behavioural* public policy? I believe it is to help us, as citizens, and as lawmakers, policy-makers, and politicians, better to understand the frailties of thought that may cloud our reasoning and judgment; and to use this understanding to inform the agreements we reach about the society in which we would like to live.

Funding

The preparation of this paper was supported by the ESRC Network for Integrated Behavioural Science (grant no. ES/P008976/1). I have benefitted enormously from discussions with many colleagues, whether in agreement or not with the approach outlined here, including Morten Christiansen, George Loewenstein, Tigran Melkonyan, Robert Sugden, and Hossam Zeitoun.

[7] Thus, in Goodman’s (1955) and Rawls’ (1971) terms, behavioural quirks can help in finding a reflective equilibrium, but the task of finding such an equilibrium remains with the people attempting to formulate scientific theories or political policies. It cannot usefully be outsourced to behaviourally informed by-standers.

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

References

Allais, M. (1953), 'Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Américaine', *Econometrica*, **21**: 503–46.

Ayton, P., A. Pott and N. Elwakili (2007), 'Affective forecasting: Why can't people predict their emotions? ', *Thinking & Reasoning*, **13**(1): 62–80.

Bacharach, M. (2006), *Beyond individual choice: Teams and frames in game theory*. Princeton: Princeton University Press.

Bernheim, B. D. and A. Rangel (2007), 'Behavioral public economics: Welfare and policy analysis with non-standard decision makers', in P. Diamond and H. Vartiainen (eds), *Economic institutions and behavioral economics*, Princeton: Princeton University Press, 7–77.

Berridge, K. C. and J. P. O'Doherty (2013), 'From experienced utility to decision utility', in P. W. Glimcher and E. Fehr (eds), *Neuroeconomics*, New York: Academic Press, 335–51.

Brickman, P., D. Coates and R. Janoff-Bulman (1978), 'Lottery winners and accident victims: Is happiness relative?', *Journal of Personality and Social Psychology*, **36**(8): 917.

Camerer, C., G. Loewenstein and M. Weber (1989), 'The curse of knowledge in economic settings: An experimental analysis', *Journal of Political Economy*, **97**(5): 1232–54.

Camerer, C., G. Loewenstein and D. Prelec (2005), 'Neuroeconomics: How neuroscience can inform economics', *Journal of Economic Literature*, **43**(1): 9–64.

Chater, N. (2018), *The mind is flat*. London: Penguin Random House/Yale University Press.

Chater, N., H. Zeitoun and T. Melkonyan (in press), 'The paradox of social interaction: Shared intentionality, we-reasoning and virtual bargaining', *Psychological Review*.

Christiansen, M. H. and N. Chater (2022), *The language game*. London, UK: Bantam Books/New York, NY: Basic Books.

Colman, A. M. (2003), 'Cooperation, psychological game theory, and limitations of rationality in social interaction', *Behavioral and Brain Sciences*, **26**(2): 139–53.

Culicover, P. W. (1999), *Syntactic nuts: Hard cases, syntactic theory, and language acquisition*. Oxford: Oxford University Press.

Edgeworth, F. Y. (1881/1961), *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. New York: Augustus M. Kelly.

Ellsberg, D. (1961), 'Risk, ambiguity, and the savage axioms', *Quarterly Journal of Economics*, **75**(4): 643–69.

Evans, N. and S. C. Levinson (2009), 'The myth of language universals: Language diversity and its importance for cognitive science', *Behavioral and Brain Sciences*, **32**(5): 429–48.

Evans, J. S. B., S. E. Newstead and R. M. Byrne (1993), *Human reasoning: The psychology of deduction*. Hove, UK: Psychology Press.

Frijters, P., A. E. Clark, C. Krekel and R. Layard (2020), 'A happy choice: Wellbeing as the goal of government', *Behavioural Public Policy*, **4**(2): 126–65.

Fukuyama, F. (2011), *The origins of political order: From prehuman times to the French Revolution*. New York: Farrar, Straus and Giroux.

Fuller, L. L. (1957), 'Positivism and fidelity to law – A reply to Professor Hart', *Harvard Law Review*, **71**(4): 630–72.

Garfinkel, H. (1967), *Studies in ethnomethodology*. Englewood Cliffs: Prentice-Hall.

Gauthier, D. (1987), *Morals by agreement*. Oxford: Clarendon Press.

Gazzaniga, M. S. (2000), 'Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition?', *Brain*, **123**(7): 1293–326.

Glimcher, P. W., M. C. Dorris and H. M. Bayer (2005), 'Physiological utility theory and the neuroeconomics of choice', *Games and Economic Behavior*, **52**(2): 213–56.

Goodman, N. (1955), *Fact, fiction, and forecast*. Cambridge: Harvard University Press.

Grice, P. (1975), 'Logic and conversation', in P. Cole and J. Morgan (eds), *Syntax and semantics 3: Speech acts*, New York: Academic Press, 41–58.

Hall, L., T. Strandberg, P. Pärnamets, A. Lind, B. Tärning and P. Johansson (2013), 'How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions', *PLoS ONE*, **8**(4): e60554. doi:10.1371/journal.pone.0060554.

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

Hammond, P. (1991), 'Interpersonal comparisons of utility: Why and how they are and should be made', in J. Elster and J. E. Roemer (eds), *Interpersonal comparisons of well-being*, Cambridge, UK: Cambridge University Press, 200–54.

Hart, H. L. A. (1958), 'Positivism and the separation of law and morals', *Harvard Law Review*, **71**(4): 593–629.

Hausman, D. M. (2012), *Preference, value, choice, and welfare*. Cambridge, UK: Cambridge University Press.

Hausman, D. M. (2015), *Valuing health: Well-being, freedom and suffering*. Oxford: Oxford University Press.

Hertwig, R., G. Barron, E. U. Weber and I. Erev (2004), 'Decisions from experience and the effect of rare events in risky choice', *Psychological Science*, **15**(8): 534–9.

Hsee, C. K. and Y. Rottenstreich (2004), 'Music, pandas, and muggers: On the affective psychology of value', *Journal of Experimental Psychology: General*, **133**(1): 23–30.

Huddleston, R. and G. Pullum (2002), *The Cambridge grammar of the English language*. Cambridge, UK: Cambridge University Press.

Hume, D. (1739/1894). 'Moral distinctions not derived from reason', In L. A. Selby-Bigge, (Ed.), *A treatise of human nature, Book III, Part I, Section I*. Oxford: Clarendon Press.

Hurst, M. (2019), *The green book: Central government guidance on appraisal and evaluation*. HM Treasury, London, UK: OGL Press.

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

Kahneman, D. (2011), *Thinking, fast and slow*. New York: Macmillan.

Kahneman, D. and A. Tversky (2000), *Choices, values, and frames*. Cambridge, UK: Cambridge University Press.

Kahneman, D., P. P. Wakker and R. Sarin (1997), 'Back to Bentham? Explorations of experienced utility', *The Quarterly Journal of Economics*, **112**(2): 375–406.

Kunar, M. A., D. G. Watson, K. Tsetsos and N. Chater (2017), 'The influence of attention on value integration', *Attention, Perception and Psychophysics*, **79**: 1615–27. doi:10.3758/s13414-017-1340-7.

Levinson, S. C. (2000), *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: MIT Press.

Lichtenstein, S. and P. Slovic (1971), 'Reversals of preference between bids and choices in gambling decisions', *Journal of Experimental Psychology*, **89**(1): 46–55.

March, J. G. and J. P. Olsen (2008), 'The logic of appropriateness', in R. E. Goodin, M. Moran, and M. Rein (eds), *The oxford handbook of public policy*. Oxford: Oxford University Press, 689–708.

Melzack, R. (1990), 'Phantom limbs and the concept of a neuromatrix', *Trends in Neurosciences*, **13**(3): 88–92.

Mercier, H. and D. Sperber (2011), 'Why do humans reason? Arguments for an argumentative theory', *Behavioral and Brain Sciences*, **34**(2): 57–74.

Mills, M., C. Rahal and E. Akimova (2020), *Face masks and coverings for the general public: Behavioural knowledge, effectiveness of cloth coverings and public messaging*. London: The Royal Society and British Academy. https://royalsociety.org/-/media/policy/projects/set-c/set-c-facemasks.pdf.

Mosteller, F. and P. Nogee (1951), 'An experimental measurement of utility', *Journal of Political Economy*, **59**(5): 371–404.

Oliver, A. (2017), 'Distinguishing between experienced utility and remembered utility', *Public Health Ethics*, **10**(2): 122–8.

Payne, J. W., J. R. Bettman and E. J. Johnson (1993), *The adaptive decision maker*. Cambridge, UK: Cambridge University Press.

Popper, K. (1972), *Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.

Rawls, J. (1971), *A theory of justice*. Cambridge: Harvard University Press.

Rizzo, M. J. and M. F. Dold (2020), 'Can a contractarian be a paternalist? The logic of James M. Buchanan's system', *Public Choice*, **183**(3): 495–507.

Scanlon, T. (2000), *What we owe to each other*. Cambridge: Belknap Press.

Sen, A. and B. A. O. Williams (Eds) (1982), *Utilitarianism and beyond*. Cambridge, UK: Cambridge University Press.

Shafir, E. (1993), 'Choosing versus rejecting: Why some options are both better and worse than others', *Memory and Cognition*, **21**(4): 546–56.

https://doi.org/10.1017/bpp.2022.2 Published online by Cambridge University Press

Shafir, E., I. Simonson and A. Tversky (1993), 'Reason-based choice', *Cognition*, **49**(1–2): 11–36.

Slovic, P. (1995), 'The construction of preference', *American Psychologist*, **50**(5): 364–71.

Stewart, N., N. Chater, H. P. Stott and S. Reimers (2003), 'Prospect relativity: how choice options influence decision under risk', *Journal of Experimental Psychology: General*, **132**(1): 23–46.

Stott, H. P. (2006), 'Cumulative prospect theory's functional menagerie', *Journal of Risk and Uncertainty*, **32**(2): 101–30.

Sugden, R. (2003), 'The logic of team reasoning', *Philosophical Explorations*, **6**(3): 165–81.

Sugden, R. (2018), *The community of advantage: A behavioural economist's defence of the market*. Oxford, UK: Oxford University Press.

Sunstein, C. R. and R. Thaler (2003), 'Libertarian paternalism is not an oxymoron', *The University of Chicago Law Review*, **70**: 1159–202.

Tobler, P. N., C. D. Fiorillo and W. Schultz (2005), 'Adaptive coding of reward value by dopamine neurons', *Science*, **307**(5715): 1642–5.

Tomasello, M. (2020), 'The moral psychology of obligation', *Behavioral and Brain Sciences*, **43**: e56. doi:10.1017/S0140525X19001742.

Tsetsos, K., N. Chater and M. Usher (2012), 'Salience driven value integration explains decision biases and preference reversal', *Proceedings of the National Academy of Sciences of the United States of America*, **109**(24): 9659–64.

Vlaev, I., N. Chater, N. Stewart and G. D. Brown (2011), 'Does the brain calculate value?', *Trends in Cognitive Sciences*, **15**(11): 546–54.

Wilson, T. D. and D. T. Gilbert (2013), 'The impact bias is alive and well', *Journal of Personality and Social Psychology*, **105**: 740–8.

Cite this article: Chater N (2024). What is the point of behavioural public policy? A contractarian approach. *Behavioural Public Policy* **8**, 197–211. https://doi.org/10.1017/bpp.2022.2