British Journal of Political Science (2022), **52**, 921–933
doi:10.1017/S0007123420000630

British Journal of Political Science

ARTICLE

Libertarian Paternalism and the Problem of Preference Architecture

Johannes Kniess

Newcastle University, Newcastle upon Tyne, UK
E-mail: Johannes.kniess@newcastle.ac.uk

(Received 6 March 2020; revised 29 June 2020; accepted 9 September 2020; first published online 25 January 2021)

Abstract

People often fail to make the choices that best satisfy their preferences. The design of the social environment inevitably makes some choices easier than others. According to Libertarian Paternalists, these facts justify governments nudging people towards better choices through changes to the so-called choice architecture. This is a form of means paternalism. However, the social environment affects not only people's choices or means, but also the preferences they adopt in the first place. Call this the problem of 'preference architecture'. This article argues that preference architecture constitutes a fundamental challenge to the justificatory basis of Libertarian Paternalism. More generally, it explores when, if ever, government paternalism that influences preference formation can be justified. While Libertarian Paternalism cannot provide a satisfactory answer, the author defends a contractualist account of paternalism based on a notion of primary goods and democratic deliberation.

Keywords: libertarian paternalism; nudges; paternalism; choice architecture; contractualism; perfectionism; preferences

It is often thought that government paternalism can only be justified, if at all, when it promotes the individual's good, as defined by his or her own preferences. However, governments often shape the social environment in ways that affect the preferences individuals adopt in the first place. On what grounds, then, can paternalism be justified?

In this article I explore this question by focusing on Libertarian Paternalism, a view that has been prominently advocated by Cass Sunstein and Richard Thaler (Sunstein 2014; Sunstein and Thaler 2003; Thaler and Sunstein 2008). Drawing on well-established findings in the cognitive sciences, the view holds that widespread or universal shortcomings in human decision making – for example, lack of self-control, the inability to deal with probability or the tendency to dismiss relevant information – should motivate governments to influence people's choices for their own good. Importantly, these paternalistic interventions are seen as most justifiable when they come in the form of 'nudges' – that is, behavioural incentives that maintain freedom of choice and impose only small costs. Nudges can range from the simple disclosure of information (for example, energy-efficiency labels for new cars) to the promotion of particular choices by making them the default option (for example, automatic enrolment in a pension plan). This is the 'libertarian' aspect of the theory – not so much because it avoids harsher means of coercion, but because people remain free to act on their own preferences and to choose their own ends.

Two main arguments are used to justify Libertarian Paternalism. The first – call it the welfarist argument – holds that governments should maximize welfare, which refers to whatever individuals believe would make their lives go well. This licences a form of government paternalism that respects people's existing preferences, particularly through nudges that preserve freedom

of choice. The second argument – call it the inevitability argument – claims that an individual's social environment inevitably structures the choices they make. Owing to this so-called choice architecture, government cannot avoid influencing individual choices in some way or another; it is 'nudging even if it does not want to' (Sunstein 2015, 6). In conjunction with the first argument, Libertarian Paternalists conclude that choice architecture should be designed to help individuals make the choices that best satisfy their preferences.

These are powerful arguments, but they are complicated by the question raised at the outset. While the inevitability argument seems to strengthen the case for government paternalism, the social environment shapes individuals' choices (or means) *and* preferences (or ends). Call this the problem of 'preference architecture'. In the first and second sections of this article, I show how this problem poses a dilemma for Libertarian Paternalists: either the inevitability argument fails to justify means paternalism, or it justifies both means *and* ends paternalism. Taking the first horn of the dilemma means that Libertarian Paternalism loses a powerful defence against sceptics, and arguably the most original justification for the theory. Taking the second horn involves abandoning libertarian commitments, and threatens to collapse the view into a form of perfectionism.

In the final section of the article, I explore the question of preference architecture in more general terms. Even if we reject the arguments put forward in favour of Libertarian Paternalism, it may nevertheless be the case that paternalistic interventions that shape future preferences are sometimes permissible. I argue that, short of subjective welfarism on the one hand, and a perfectionist account of the good on the other, we can maintain a moderate commitment to respect people's preferences through a contractualist account that models generic interests in pluralistic societies. Government paternalism can be guided by an account of goods that any rational and reasonable person desires. Because the promotion of some of these goods can conflict with that of others, however, the theory has to be complemented by an account of free and open democratic deliberation on the priorities of preference architecture. This approach offers a justifiable basis for paternalistic policies that are not perfectionist in character. This way, then, we can reconcile the fact that the social environment shapes our preferences with a deep-seated scepticism about government paternalism that disregards our own ideas of the good life.

The Case for Libertarian Paternalism

Paternalism is conventionally understood as interfering with a person's liberty or autonomy 'for his own good' (Dworkin 1972, 67). To many people it is a pejorative term: the fact that a policy or institution is paternalistic is a reason to reject it. At the same time, however, many paternalistic policies seem largely uncontroversial. Few people object to laws forcing drivers to wear seat belts, or food hygiene regulations that keep dangerous products off the shelves. To resolve this seeming tension, it is helpful to treat paternalism as a non-evaluative term, and to draw a number of distinctions that track the conditions under which it may be justifiable. Two of these are particularly important to understand Libertarian Paternalism.

The first distinction concerns means and ends. Ends paternalists seek to guide people towards goals that they do not identify as their own. For example, they might promote sexual abstinence policies even when the supposed beneficiaries do not agree that abstinence is a worthwhile goal. Means paternalists, by contrast, respect people's preferences or ends, but intervene when the means they use to pursue them are misguided by some form of non-voluntariness or irrationality. Imagine a person who is about to drink poison, falsely believing it is wine. Means paternalists would interfere because the person does not want to poison herself. Even John Stuart Mill, who is commonly credited with formulating the classic liberal rejection of paternalism, argued that it was justified to interfere with actions that would endanger the life of a person who is 'delirious, or in some state of excitement or absorption incompatible with the full use of the reflecting faculty' (Mill 1991[1859], 107). Libertarian Paternalists stand in the tradition of means paternalism, but considerably expand the scope of permissible interventions. For, as well-known findings

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

in the cognitive sciences show, non-voluntary or irrational acts are not all that rare: rather, they are widespread or universal features of our decision making.[1] This, in principle, opens the door to considerably more paternalistic interventions than Mill, for example, had in mind.

The second distinction tracks the degree of invasiveness of the interference: soft interventions push people to make one choice over another, while hard or coercive efforts restrict their freedom by taking away choices.[2] Requiring mountaineers to take out insurance before they climb a dangerous mountain is an example of the former approach; forcibly stopping them from climbing it is an example of the latter. There is some discussion about whether interventions that do not take away any choices – say, attempts to provide people with information about the risks of mountaineering – are paternalistic at all (Hausman and Welch 2010; Lichtenberg 2016). However, it makes sense to follow Thaler and Sunstein in understanding paternalistic interventions as a *continuum*: whereas hard approaches impose large costs on those who reject the intervention, soft paternalism involves smaller (often immaterial) costs (Sunstein and Thaler 2003, 1,185). Paternalistic policies can fall anywhere between these extremes.

Libertarian paternalism typically advocates soft approaches. Nudges, which are designed to alter behaviour without forbidding any options are the paradigmatic case. But it is worth noting that nudges and Libertarian Paternalism are not coextensive (Saghai 2013, 488). For one thing, some nudges are not paternalistic. Changing the default option for organ donations from 'opt-in' to 'opt-out' is a nudge that can save many lives, but not that of the organ donor herself. Hence this is not an example of an intervention designed to promote the person's own good. Secondly, Libertarian Paternalists prefer nudges, but they do not rule out harder forms of paternalism when 'strong empirical justifications, involving relevant costs and benefits, support a more aggressive approach' (Sunstein 2014, 72). This suggests that it is not the *softness*, but rather the non-interference with people's ends that gives the theory its libertarian credentials. In other words, Libertarian Paternalists are means paternalists first and foremost, and only then soft paternalists.

How do proponents of Libertarian Paternalism attempt to justify it? In most of their writings, Thaler and Sunstein motivate their theory by providing empirically informed – and often very persuasive – examples of nudges that most people would be happy to turn into legislation. Though they never quite present it as a systematic theory, Sunstein's book *Why Nudge?* sets out the most comprehensive defence of the normative foundations of Libertarian Paternalism. Two arguments lie at its core: the welfarist and inevitability arguments. I next examine each in turn and explain how they are jointly thought to provide a justification for the theory.

> *Welfarist Argument*. The 'master concept' of government action is the promotion of social welfare, which refers to the aggregation of individual welfare (Sunstein 2014, 13). Individual welfare, in turn, is simply 'whatever [individuals] think would make their lives go well' (13). In other words, government should 'respect people's ends' (138). Because of behavioural biases, misinformation or weakness of will, however, people often fail to advance their own ends. In order to promote welfare, then, government may engage in means (but not ends) paternalism.

This is a familiar argument in favour of means paternalism. It states merely that governments should be paternalistic when it improves people's welfare. Now for a variety of reasons – including the fallibility of government officials – paternalistic interventions may sometimes inadvertently *decrease* welfare. But this cannot justify a blanket prohibition on paternalism, according

[1]See, e.g., Kahneman (2011). Hence, the subjective welfarism that Libertarian Paternalists embrace is one based on 'informed' or 'reconstructed' preferences, that is, preferences people would have if they had 'complete information, unlimited cognitive abilities, and no lack of self-control' (Sunstein and Thaler 2003, 1,162).

[2]I follow Sunstein's use of the terms 'soft' and 'hard'. According to the standard account in the literature (e.g., Feinberg 1986), soft paternalism refers to interferences that target subjects who do not act voluntarily or autonomously. Understood in this latter sense, soft paternalism is a version of means paternalism.

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

to Sunstein, because worries that paternalism sometimes reduces welfare 'have no force when some kind of paternalism is inevitable' (2014, 121). Hence the first argument is particularly salient when supported by the second, more original, argument.

> *Inevitability Argument*. Government has to make decisions that often inevitably affect the social environment in ways that will influence people's choices. Because it is not possible to dispense with a social environment, 'choice architecture is inevitable' (Sunstein 2014, 118–9). Therefore it is not impermissible for government to influence people's choices.

A classic example of choice architecture is the 'cafeteria case', in which Sunstein and Thaler imagine a cafeteria manager who must decide how the food will be displayed and arranged. If customers have to walk past the salad bar to get to the burger stand, for example, they are more likely to purchase a salad. Yet if burgers are the first thing they see, they might quickly forget about the salad. Of course, the food has to be displayed somehow, and influencing customers' choices is therefore unavoidable. If that is the case, the best choices should be promoted.[3] 'In an important respect the anti-paternalist position is incoherent', Sunstein and Thaler conclude, 'simply because there is no way to avoid effects on behavior and choices' (Sunstein and Thaler 2003, 1,182; see also Sunstein 2014, 121).

At first sight, this conclusion may seem like a non sequitur. After all, the inevitability argument only shows that influencing choices is unavoidable and therefore not impermissible; it does not specify the ways in which we may permissibly do so. Rather than implementing a paternalistic choice architecture, government could also design the social environment in ways that promote non-paternalistic ends – that is, ends that are not meant to directly benefit the intended target for his or her own good. These might include interventions to avoid harm to others, provide public goods or promote other just ends.[4] Furthermore, one might object that there is an important difference between intentionally and unintentionally influencing people's choices: since the former imposes the will of one agent on another, it may face a higher justificatory burden (Hausman and Welch 2010, 133; Schmidt 2017). In what sense, then, does the inevitability argument support government paternalism?

For Libertarian Paternalists, the answer must lie in the link between the welfarist and inevitability arguments. Choice architecture is unavoidable, and government is always influencing individual choices – whether by default or by design. There is a sense in which, once we acknowledge this, all nudges are intentional. It is true that this insight alone cannot justify paternalism. But if we agree that the role of government is to promote welfare, and that welfare can be improved through paternalistic measures, then the inevitability argument implies a strong role for paternalistic interventions. In other words, if we accept the inevitability argument *in conjunction with* the welfarist argument, then one might argue that the burden of proof is on those who reject influencing people's choices in order to improve their welfare – as long as the latter is understood in a way that reflects people's own ideas of what makes their lives go well.

The Problem Of Preference Architecture

Libertarian Paternalism has attracted a great deal of criticism, ranging from concerns about lacking transparency, manipulation, condescension and disrespect, to worries about its

[3]Sunstein and Thaler argue that the healthier options should be displayed first, given that most people want to be healthy. Of course, patrons may have different preferences, and at least some are likely to genuinely privilege tasty food over their own health. Hence the nudge that Sunstein and Thaler propose would be a case of 'impure paternalism', in which the class of persons whose choices are affected is larger than that for which the interference is justified on the basis of their own good. This is not a problem specific to Libertarian Paternalism but rather a general feature of government paternalism, which typically affects large numbers of people simultaneously. Cf. Mills (2018), 400.

[4]Consider, for example, 'green' nudges that aim to instil environmentally conscious behaviours. For more detailed discussions of non-paternalist nudging, see Kelly (2013); Moles (2015).

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

implementation and potential abuses (Grüne-Yanoff 2012; Hausman and Welch 2010; Rebonato 2012; Waldron 2014; Wilkinson 2013).[5] I believe some of these worries are warranted, and that Libertarian Paternalists need to do further work to defend their agenda. However, in this article I want to bracket these concerns, as far as possible, in order to focus on an unresolved challenge that has received very little attention: the implications of the inevitability argument on the distinction between means and ends paternalism. To develop the point, take the common observation that the social environment does not only influence people's choices or means, but also their preferences and decisions about which ends to pursue in the first place. This is by no means a new or controversial claim. When discussing the design of the basic institutional structure of society, for example, John Rawls points out that:

> The social system shapes the wants and aspirations that its citizens come to have… Thus an economic system is not only an institutional device for satisfying existing wants and needs but a way of creating and fashioning wants in the future… These matters are, of course, perfectly obvious and they have always been recognized. They were stressed by economists as different as Marshall and Marx (Rawls 1971, 259; see also Barry 1965, 77).

This notwithstanding, contemporary economists tend to assume that individuals have stable and exogenous preferences.[6] Much like perfect rationality or pure self-interest, however, this assumption is a methodological abstraction that only partially captures human behaviour. Indeed, the design of the social environment, including its economic institutions, shapes individual preferences through a broad range of mechanisms: by affecting individual motivations, influencing the evolution of norms or triggering cultural learning processes (Aaron 1994; Bowles 1998).

It is not hard to see how some of the policies proposed by Libertarian Paternalists can contribute to these processes. Consider, for instance, default rules. Evidence suggests that people often understand default rules – such as those concerning enrolment in pension plans – as implicit advice, or as an indication of the popularity of a choice (Willis 2013, 1,168–9). Preferences might then be influenced by a desire to conform to perceived social expectations or norms. Default rules can also shape preferences by enabling experiences; in other words, a decision maker might 'try it and like it' (Willis 2013, 1,169). The 'mere-exposure effect', which is well documented in the social psychological literature, can lead people to develop preferences and positive affective reactions through repeated exposure to a given stimulus (Zajonc 2006).

Suppose, then, that there are cases in which governments have to make inevitable decisions about how to arrange the social environment; these decisions influence or alter people's preferences or ends. Recall the cafeteria case. Seen as a single instance of a nudge, it may well be true that the design of the cafeteria line influences only the choices that customers make, and not their fundamental preferences. No one will lose their passion for burgers because one day they were hidden away in the corner of the canteen. But imagine that the new design becomes permanent by law, and that some people never get to experience a different kind of cafeteria. Over long periods of time, preferences and tastes are likely to be influenced by the way options are presented. Just as we have said that 'choice architecture' is frequently inevitable, we might conclude that 'preference architecture' is also inevitable.

> *Modified Inevitability Argument*. Government has to make decisions that often inevitably affect the social environment in ways that will influence people's preferences. Because it is

[5]For replies to some of the main lines of criticism, see Sunstein (2015).

[6]Gary Becker's (1976, 14) statement that 'all human behaviour can be viewed as involving participants who maximize their utility from a stable set of preferences' is representative in this regard. Becker later revised his position, arguing that 'modern economics has lost a lot by completely abandoning the classical concern with the effects of the economy on preferences and attitudes' (1996, 19).

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

> not possible to dispense with a social environment, 'preference architecture' is inevitable. When this is the case, it is not impermissible for government to influence people's preferences.

Note that the argument does not claim that the social environment shapes all preferences, or that its influence is inescapable for everyone. It merely shows that whenever influencing preferences is unavoidable, it is not impermissible to do so. As such, it is a logical extension of the inevitability argument. Now, if Libertarian Paternalists use the latter to justify soft means paternalism, then by parity of reasoning they must also accept some form of soft *ends* paternalism. For, in conjunction with the welfarist argument, the modified inevitability argument seems to suggest that governments are justified in influencing people's preferences for their own good. But this not only flies in the face of the Libertarian Paternalist commitment to means (not ends) paternalism. Perhaps even more worryingly, it also seems to cast doubt on the subjective definition of welfare on which the welfarist argument is based.[7]

Thaler and Sunstein are not unaware of these issues. In the cafeteria case, they discuss the seeming impossibility of arranging the food in the way patrons would have chosen on their own. 'If the arrangement of the alternatives has a significant effect on the selections the customers make', they note, 'then their true 'preferences' do not formally exist' (Sunstein and Thaler 2003, 1,164). Hence the cafeteria case is not simply one of helping people overcome their weakness of will in order to satisfy a previously formed preference for healthy food. Rather, in this case the paternalist is *creating* a preference for healthy food where there might not have been one before. As Sunstein puts it elsewhere, 'choice architects might be engineering the very judgment from which they are claiming authority' (Sunstein 2015, 21). In another example from *Why Nudge*, Sunstein discusses government-mandated fuel economy labels for new vehicles. Here too, he admits that even the softest of nudges – the disclosure of information – can influence both means and ends because the highlighted information is necessarily selective: it informs us about one attribute that people might value in a new car (energy efficiency) but not others (speed, design and so on) (Sunstein 2014, 69). This in turn is likely to affect people's views about what they should value in cars: they may come to think that they like hybrid Hyundais more than gas-guzzling Hummers. Sunstein, however, is not particularly troubled by this:

> We really should not be too fussy or clever here, and we should avoid tying ourselves into conceptual knots. If framing or selectivity is at work, there may be a form of ends paternalism, but it is likely to be of a modest kind, perhaps so modest that we do not have to worry much (Sunstein 2014, 69).

Unfortunately, it is unclear in what sense the risk of ends paternalism is 'modest'. At one point, Sunstein observes that if paternalistic interventions promote preferences that people already cared about before, 'then it is fair to say that [this] paternalism is at least centrally about means, and that the intrusion on people's ends is small and possibly even incidental' (Sunstein 2014, 69). But that, of course, is trivially true: if governments promote ends that people already endorse, they are not changing their ends. The problem is that people may have chosen different ends, or would have prioritized their ends differently to the way the paternalist envisages.

Perhaps by 'modest', Sunstein means that the ends shaped by preference architecture are not particularly important or controversial ones. This is suggested by the choice of examples: our

[7]It may also heighten some of the other worries commonly raised in the literature, such as concerns about disrespect. Some argue that because nudging targets means rather than ends, 'it does not involve a negative judgment on a particular agent' (Moles 2015, 654). This kind of defence may not available to proponents of Libertarian Paternalism if their policies do indeed affect the ends that agents come to have. The problem of preference architecture may also exacerbate worries about manipulation, if the latter is understood as a form of non-rational influence that 'threatens peoples' ability to (…) form and pursue their ends independently' (Moles 2015, 654).

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

taste in cars is perhaps a rather banal example of an individual preference, and maybe nobody minds being softly nudged towards having a salad instead of a hamburger. Yet there are two problems with this interpretation. First, if we want to respect people's views about what matters in life, who is to say which ends are (or are not) important? Secondly, and perhaps more importantly, there is no reason to suppose that preference architecture does not affect all kinds of ends. Some may be uncontroversial, others less so. Take, for instance, religious education in public schools. Whether children (or their parents) opt in or out of religious education classes is a question that is as controversial as it is unavoidable.

Consider, finally, a third interpretation of what Sunstein might mean by 'modest'. Perhaps the point is that preference architecture has merely a weak impact on people's behaviour. Since, by definition, 'soft' interventions impose only small costs on those who refuse to change their mind, one could think that they will only have a weak effect. Of course, in this general form, this kind of reply would contradict the empirical evidence on behavioural biases that undergirds the case for Libertarian Paternalism. The theory is motivated by, and clearly depends on, evidence that people *do* respond to framing effects built into the environment. Hence the point would have to be that the environment influences our choices, but not our preferences. Some means paternalists endorse this claim. Sarah Conly (2012, 124), for example, asserts that most of our deeply held ends are relatively stable, and that, unlike our means, they 'do not change according to the peculiarities of choice procedure'.

In response, it is useful to take a closer look at the distinction between means (or choices) and ends (or preferences). Although we have so far assumed that both categories can be clearly demarcated, the distinction is often blurry. Some ends are, in fact, means to more 'final' ends. For example, a student might have the goal of graduating from medical school. Obtaining her degree is an end, but also a means to another end, such as a successful career as a doctor. And that end might in turn be a means to an even more fundamental end, such as social recognition or a desire to help others. So rather than a binary distinction between means and ends, we may think instead of a hierarchy of ends. While some ends are more immediate or instrumental, others are 'final' in a way that gives special depth and meaning to the narratives of our lives.

Consider now a second distinction. Some means and ends are particularly malleable, or endogenous to the setting in which they are expressed. Others are more robust, in the sense that people do not change them easily. Now, Conly's claim is that our more final ends – those higher up in the hierarchy of ends, as it were – are robust in a way that our means and our more instrumental ends are not. They do not change depending on the preference architecture that happens to be in place. One could interpret evidence in the cognitive sciences in a way that would seem to support this claim. Psychologists and behavioural economists who have studied the phenomenon of preference reversals, for example, have done so primarily in experimental settings where subjects are presented with unfamiliar, low-stakes choices (Lichtenstein and Slovic 2006, section II). In these experimental settings, people often change their preferences about, say, their willingness to pay for some consumer product. But they are unlikely to change their preferences about deeply held convictions. A conservative Christian, for example, might not change her views on abortion simply by being asked the same question in different ways.

But, of course, a narrow focus on experimental evidence does not show that our more final ends are robust, at least not in the sense that they are formed independently of the social environment. If we take a broader view of how life is conditioned by the social world that surrounds us, it is hard to deny that government regulation in areas such as work, marriage, child rearing or education profoundly shapes our goals, beliefs, desires and aspirations. Even our political preferences, such as support for democracy, are strongly influenced by the political systems that are in place (Fuchs-Schündeln and Schündeln 2015). Indeed, as Brian Barry once remarked, 'the dependence of wants on the social environment is a commonplace of sociology, which could almost be defined as the intellectual consequences of taking it seriously' (Barry 1965, 75).

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

Hence, even if it is true that our more immediate or instrumental ends are *particularly* responsive to preference architecture, in the long view our more final ends are also susceptible to its effects.

And so the problem for Libertarian Paternalism is the following. If we accept the implications of the inevitability argument, as I think we must, then government cannot avoid influencing both our choices and our preferences. If the welfarist argument provides a rationale for paternalism, then in conjunction with the inevitability argument it would seem to justify both means and ends paternalism. This creates a dilemma. Libertarian Paternalists could affirm the permissibility of ends paternalism. This option entails abandoning subjective welfarism: government cannot advance people's good by overriding their preferences if their good is defined in terms of these preferences. Indeed, in some of his recent writings, Sunstein has hinted that people's views about what would make their lives go well are not always authoritative, and that paternalists must sometimes make 'independent' judgements about welfare (2019, 102). But accepting this conclusion seriously undermines the Libertarian Paternalist rejection of ends paternalism, and thus the 'libertarian' credentials of the theory.

The more attractive option for Libertarian Paternalists would be to maintain that the inevitability of preference architecture does not justify ends paternalism. This option, however, comes at the cost of conceding that the inevitability of choice architecture does not justify means paternalism either. To be sure, government cannot avoid influencing people's choices and preferences when it designs the social environment. But if the lodestar is not always welfare, then the social environment need not be designed to increase welfare through paternalistic interventions. In other words, the problem of preference architecture shows that the link between the inevitability argument and the welfarist argument does not hold water. The view therefore loses one of its most powerful and original arguments.

The Permissibility of Shaping Preferences

Going beyond Libertarian Paternalism, a difficult question raised by the problem of preference architecture remains. While the fact that government is constantly shaping the social environment is not a *justification* for paternalism, it may well be a permissible *instrument* to make people better off in paternalistic ways. Let us assume, for the purposes of this article, that the evidence on widespread errors in human decision making does indeed sometimes lend support to paternalistic interventions. Furthermore, let us assume that ends paternalism faces a higher justificatory burden than means paternalism. After all, most of us form our preferences and pursue our ends believing that these give meaning to our lives precisely because they are *our* preferences and *our* ends. Given these assumptions, the problem can be put as follows: even if government is justified in designing paternalistic nudges that respect people's own ideas of the good – in other words, if means paternalism is permissible – the problem of preference architecture complicates the task. For one thing, what passes as means paternalism in a single iteration may shape ends over long periods of time. In the cafeteria case, for example, repeated nudges towards healthy food may change or cultivate preferences, rather than simply help satisfy existing wants. Even more fundamentally, robust preferences might not formally exist before the social environment is designed. Thus when some form of means paternalism is in principle justifiable, but government cannot help influencing people's ends, we require an account of the ways in which government can permissibly influence ends. While Libertarian Paternalism does not provide such an account, I briefly sketch some possibilities in the remainder of the article.

The first is simply a modified version of subjective welfarism. Despite the difficulties described above, one could insist that governments help people pursue the ends they already hold. A proponent of this view might argue as follows: 'Granted, some of our more immediate or instrumental ends may not be stable enough to serve as the basis for paternalistic intervention. But government paternalism could focus at least on those ends that we hold in a relatively robust way. Now, of course, even our more robust ends may have been formed in light of the current

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

way the social environment is structured, and through paternalistic interference, government may be further reinforcing those ends. But this is merely an unintended side effect of these interferences, and after all, we have to start somewhere. The fact that our ends are shaped by the social environment does not render them inauthentic, or else there would be no authentic ends. Hence, paternalistic interventions should always be guided by people's existing ends.'

By focusing narrowly on the preferences that people have at a given time, this approach ignores the preference-forming aspect of means paternalism. But this, it seems to me, is a mistake. When paternalistic interventions shape or consolidate preferences in the future, they seem to call for a more demanding justification. This is so because government is not simply helping individuals attain their own ends; it is also actively choosing to influence people's future preferences. If paternalism were unavoidable, one could perhaps brush aside this concern as the unintended side effect of *any* government action. But, as we have seen, it is not. Hence paternalistic preference architecture needs to be held to a higher justificatory bar.

This becomes particularly evident when we focus on the problem of adaptive preferences (Elster 1983; Nussbaum 2000, Chap. 2). Existing preferences are shaped by the social environment, and that environment is often unjust. By respecting and promoting these preferences, government paternalism therefore risks condoning and reproducing an unjust status quo. This is a point that Sunstein himself developed in some of his earlier work (Sunstein 1991). Taking examples such as women who internalize unfair traditional gender roles in order to avoid stigma, Sunstein argues that democratic governments should not be concerned merely with satisfying the preferences that people already have, but rather override preferences when doing so would promote welfare or autonomy. In stark contrast with his work on Libertarian Paternalism, he thus concludes that 'subjective welfarism, even as a political conception, is unsupportable' (Sunstein 1991, 8).

Now, one might argue that even subjective welfarism can be circumscribed by independent moral principles that rule out the promotion of 'unjust' preferences, such as those based on sexist societal expectations. This would require a theory of justice to identify preferences that are unsuitable goals for preference architects. But leaving this issue to one side for the moment, the reproduction of existing preferences through paternalistic interventions can be problematic even when these preferences are not unjust. Imagine a society where, thanks to widespread cultural norms, many or most people develop a preference for an unhealthy diet. Suppose further that some people occasionally fail to act on that preference. Through (means) paternalistic interventions, government nudges these people towards their unhealthy diet, thereby helping to reproduce the cultural norms on which the preference is based. Even if any given individual were happy to be nudged in this way, it is not clear that satisfying current individual preferences justifies the government's role in shaping future preferences. In this case, at least, it seems sensible to maintain that government should not promote existing preferences. But on what grounds can such a form of preference architecture be rejected?

According to one influential view in the history of moral and political thought, shaping individual preferences in desirable ways is one of the central functions of government. *Perfectionism*, though a broad family of views, typically describes the idea that governments should promote certain conditions for human flourishing or the good life (for example, Chan 2000, 5). The good life is not defined by people's actual preferences or attitudes, but rather in reference to certain objective goods or values, such as health and longevity. Thus a perfectionist government would design preference architecture so as to encourage people to adopt certain objectively valuable ends, which they may or may not view as valuable.

Perfectionism gives rise to several concerns that have preoccupied critics. Perhaps most importantly, there is considerable disagreement about what the good life is. Even if one accepts a form of perfectionism that is pluralist in recognizing several reasonable paths to flourishing, there is disagreement about which paths those are. Hence government might lack political legitimacy if it enforces some paternalistic intervention by appealing to reasons that not all reasonable

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

people can accept (Quong 2011; Rawls 1993). Furthermore, if governments attach no significance to people's own attitudes towards the good – even when these happen to coincide with the official view – they could be seen to express a worrying lack of respect for their citizens. Perfectionism also raises a number of somewhat more practical worries related, for instance, to its inherent political instability in a pluralistic society, or to the risk of entrusting the promotion of the good to fallible, corruptible and potentially repressive governments.

Of course, none of these worries may prove to be decisive. Liberal Perfectionists have put forward sophisticated accounts that address some of these worries, in particular by championing non-coercive policies and putting individual autonomy at the centre of their accounts of the good life (Raz 1986; Wall 1998). Fully discussing the merits of these approaches is beyond the scope of this article. Instead, I want to show that perfectionism is not the only plausible answer to the problem of preference architecture. It is possible for government to enact paternalistic interventions that shape future preferences without making authoritative judgments about what makes a life go well.

On a *contractualist* approach, any exercise of government power ought to be guided by principles that we could justify to one another under conditions that reflect our status as moral equals and our capacity to act as rational agents. This idea, of course, finds its most prominent expression in Rawls, who outlines an 'original position' in which parties to a hypothetical social contract decide on principles of justice without knowledge of their own particular conceptions of the good (Rawls 1971). Behind the 'veil of ignorance', it is rational for parties to agree to paternalistic principles to protect themselves from serious shortcomings in their decision making. Like Odysseus commanding his sailors to tie him to the mast, citizens would agree to a kind of insurance policy that authorizes government to safeguard their interests when their judgment is impaired (Dworkin 1972: 77–8; Rawls 1971, 248–50). But because contract parties do not know their own ends, they require some measure of the interests or expectations of representative people. Given certain basic assumptions about human psychology and the social sciences, we can thus identify a catalogue of 'primary goods' that any rational person would want in societies like ours – 'things that men are presumed to want whatever else they want.' (Rawls 1971, 260). The list would certainly include items like health and wealth, but more generally any good that could serve as an all-purpose means for people to construct their individual life plans and pursue their ends. To be sure, primary goods are not the only goods that rational people might want. Some goods – say, a religious education – are valuable only for the pursuit of certain life plans. Other goods may be valuable for the pursuit of ends that are *unreasonable*, that is, incompatible with fair institutional arrangements that we can justify to one another. In contrast, the notion of primary goods can serve as a generic account of the typical interests that people in a pluralistic society have.

How can a contractualist approach, with its corresponding account of primary goods, help resolve the problem of paternalistic preference architecture? Here it is useful to recall the distinction between robustly held preferences and more malleable preferences. If a person's preferences are malleable, or endogenous to the social environment in which they are formed, the account I have sketched implies that preference architecture should be designed so as to improve the individual's share of primary goods (see also Rawls 1971, 249). This might justify, for instance, interventions that extend longevity or ensure sufficient savings for old age. The pursuit of these ends need not be based on perfectionist standards.[8] Rather, by mirroring the position of rational contract parties, these interventions simply ensure the possession of goods that would be useful for any reasonable life plan. What if a person's preferences are not malleable, but robust? Here, government should use these preferences as a guide for paternalistic interventions

[8]It is often wrongly assumed that, short of subjective welfare, this is the only way to justify nudges. See, e.g., Blumenthal-Barby (2013, 180): 'there is a sense of an underlying perfectionist standard of good, namely health and wealth, to which the various nudges direct the masses'.

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

under only two conditions: (1) they are reasonable (that is, they do not undermine the fairness of basic social arrangements)[9] and (2) helping to satisfy these preferences does not shape future preferences in a way that undermines people's shares of primary goods. Recall the example of a society where widespread cultural norms create preferences for an unhealthy diet. To the extent that health is a primary good, the logic of a paternalistic insurance policy issued behind the veil of ignorance suggests that preference architecture should not be put in place to promote these preferences.

This, in broad terms, is the contractualist solution to the problem of preference architecture. But there is an important complication. While every single primary good is plausibly desired by the parties in the original position, in practice, different individuals may desire different combinations of primary goods. And the promotion of these goods can conflict. Consider, for instance, the relationship between health and wealth. In areas such as environmental regulation or workplace health standards, governments are often confronted with difficult trade-offs. The starkly diverging attitudes towards the regulation of new technologies in the European Union and the United States, for instance, reflect regulatory cultures that either emphasize innovation and economic growth, or take a precautionary attitude to possible health hazards. Similarly, conflicting incentives in the area of labour law encourage the formation of preferences for consumption on one side of the Atlantic, and preferences for leisure time on the other. The account of primary goods in itself cannot resolve these trade-offs because it is rational to want all of these goods. As critics have pointed out with regard to Rawls's theory of justice, this creates an 'indexing problem' when comparing different individual bundles of goods to measure who is more or less disadvantaged in society (Gibbard 1979). But it also gives rise to a similar problem when using primary goods as a guide for paternalistic preference architecture.

I argue that in both cases there is only one plausible solution: societies must ultimately establish their priorities via democratic deliberation. In other words, when the approach described here reaches its limits, the veil of ignorance must be lifted to allow real citizens to collectively decide how trade-offs among disparate goods and social goals ought to be resolved (see also Kniess 2019, 418–20). The account of primary goods merely sets the parameters of acceptable paternalistic interference; the precise content is a matter for politics. On the face of it this argument may appear to be somewhat circular. Preference architecture is supposed to guide the development of citizen's preferences, yet it is itself (at least in part) the outcome of collective preferences expressed in the political domain. However, democratic deliberation does not only confer legitimacy on the government's power to shape future preferences; under conditions of free and reasoned public discussion, existing preferences are also subject to change.

The open-ended nature of this approach is a strength rather than a weakness. Libertarian Paternalism, despite its insistence on respecting people's subjective preferences, often arouses suspicions related to its apolitical and technocratic tendencies. 'A large virtue of technocrats in government' Sunstein maintains, 'is that they can help overcome some of the errors that might otherwise influence public as well as private judgments' (Sunstein 2014, 101). Critics have seen herein a version of what Bernard Williams called 'government house utilitarianism', the idea that an enlightened elite formulates simple rules for common people to follow, believing that the full complexity of moral decision making would overwhelm them (Waldron 2014, citing Williams 1985). The worry stands even if, as Libertarian Paternalists insist, government communicates their policy goals transparently and publicly (Thaler and Sunstein 2008, 244). After all, a policy may be suitably transparent but not democratically legitimate (Schmidt 2017, 411). A

[9]Imagine, for example, that owing to internalized racism, members of a disadvantaged minority group have a robust preference to send their children to de facto segregated schools. If, for whatever reason, members of this group fail to act on this preference, should the government nudge them towards it? Surely not: in order to satisfy principles of justice that citizens would choose behind the veil of ignorance (principles that would include safeguards for fair educational opportunities), preferences like these cannot serve as the basis of paternalistic intervention.

https://doi.org/10.1017/S0007123420000630 Published online by Cambridge University Press

deliberative process, by contrast, expresses mutual respect by treating all participants as equals in providing reasons for collective decision making in the public sphere. Rather than leaving decisions about preference architecture to technocrats behind the closed doors of government, democratic deliberation is, by its very nature, participative and open ended.

All this, I am aware, requires considerably more detail. But if the approach that I have sketched here can be fleshed out, it offers a plausible solution to the problem of preference architecture. If some form of means paternalism is in principle justifiable, we require an account of the ways in which government, as a result of helping individuals achieve their ends, may permissibly shape people's future preferences. Short of subjective welfarism on the one hand, and a perfectionist account of the good on the other, a contractualist approach to government paternalism suggests that future preference formation needs to be guided by principles that command the support of reasonable and rational people. The notion of primary goods and democratic deliberation can thus give content to the contours of paternalistic preference architecture.

Conclusion

Among the most interesting, yet least discussed, aspects of Libertarian Paternalism is the argument concerning the inevitability of government intervention. We do not make choices in a vacuum, after all, and the social background against which we choose must be designed in one way or another. This fact, combined with a commitment to increasing welfare, is thought to support government paternalism. It is a powerful argument, but the implications are far-reaching. For although it is true that the social environment shapes individual choices, the same holds for preferences. If one thinks that means paternalism is justified by the inevitability of choice architecture, one must also be committed to the view that ends paternalism is justified by the inevitability of preference architecture. This leads to a dilemma.

If, on the one hand, one denies that the inevitability of choice or preference architecture justifies paternalism of any kind, then Libertarian Paternalism will have lost one of its most compelling motivating arguments. It will also require some account of what exactly government is justified in doing when some form of nudging is inevitable. If, on the other hand, one concedes that the inevitability of preference architecture supports some form of ends paternalism, then Libertarian Paternalism can no longer maintain a focus on means paternalism *or* subjective welfarism.

In addition to this negative conclusion, I also advanced a positive suggestion. Even if we reject Libertarian Paternalism, we still require a normative account of the permissibility of preference architecture. Governments can often design the social environment in ways that advance non-paternalistic goals, and often they should. But at least in some cases, a paternalistic preference architecture may also be justifiable. I have argued that this need not presuppose a perfectionist account of the good. Instead, a contractualist account of paternalism can provide us with a list of generic goods that rational and reasonable people in pluralistic societies desire. Since the promotion of these goods can conflict with each other, however, there must also be room for democratic mechanisms to weigh these goods against each other. Hence, while the inevitability of preference architecture is not a direct justification for paternalism, paternalistic interventions that shape people's future preferences can nonetheless be justified on independent grounds.

Acknowledgements. I thank Harrison Frye, Suzie Kim, David Miller, Andrea Sangiovanni and audiences in Pavia, Harvard and the European University Institute for helpful feedback on earlier drafts. I am particularly grateful to Julian Urrutia, who helped me think about and develop some of the central ideas of this article.

References

Aaron H (1994) Public policy, values, and consciousness. *The Journal of Economic Perspectives* **8**(2), 3–21.
Barry B (1965) *Political Argument*. Berkeley and Los Angeles: University of California Press.

Becker G (1976) *The Economic Approach to Human Behavior*. Chicago, IL: Chicago University Press.
Becker G (1996) *Accounting for Tastes*. Cambridge, MA: Harvard University Press.
Blumenthal-Barby JS (2013) Choice architecture: a mechanism for improving decisions while preserving liberty? In Coons C and Weber M (eds), *Paternalism: Theory and Practice*. Cambridge: Cambridge University Press, pp. 178–196.
Bowles S (1998) Endogenous preferences: the cultural consequences of markets and other economic institutions. *Journal of Economic Literature* **36**(1), 75–111.
Chan J (2000) Legitimacy, unanimity and perfectionism. *Philosophy & Public Affairs* **29**(1), 5–42.
Conly S (2012) *Against Autonomy: Justifying Coercive Paternalism*. Cambridge, MA: Cambridge University Press.
Dworkin G (1972) Paternalism. *The Monist* **56**, 64–84.
Elster J (1983) *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge, MA: Cambridge University Press.
Feinberg J (1986) *Harm to Self*. Oxford: Oxford University Press.
Fuchs-Schündeln N and Schündeln M (2015) On the endogeneity of political preferences: evidence from individual experience with democracy. *Science* **347**(6226), 1145–1148.
Gibbard A (1979) Disparate goods and Rawls' difference principle: a social choice theoretic treatment. *Theory and Decision* **11**(3), 267–288.
Grüne-Yanoff T (2012) Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare* **38**(4), 635–645.
Hausman D and Welch B (2010) Debate: to nudge or not to nudge. *The Journal of Political Philosophy* **18**(1), 123–136.
Kahneman D (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
Kelly J (2013) Libertarian paternalism, utilitarianism, and justice. In Coons C and Weber M (eds), *Paternalism: Theory and Practice*. Cambridge: Cambridge University Press, pp. 216–230.
Kniess J (2019) Justice in the social distribution of health. *Social Theory and Practice* **45**(3), 397–425.
Lichtenberg J (2016) For your own good: informing, nudging, coercing. *Georgetown Journal of Law and Public Policy* **14**, 663–682.
Lichtenstein S and Slovic P (eds) 2006. *The Construction of Preference*. Cambridge: Cambridge University Press.
Mill JS (1991 [1859]) *On Liberty and Other Essays*. Oxford and New York: Oxford University Press.
Mills C (2018) The choice architect's trilemma. *Res Publica* **24**(3), 395–414.
Moles A (2015) Nudging for liberals. *Social Theory and Practice* **41**(4), 644–667.
Nussbaum M (2000) *Women and Human Development*. Cambridge, MA: Cambridge University Press.
Quong J (2011) *Liberalism Without Perfection*. Oxford: Oxford University Press.
Rawls J (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls J (1993) *Political Liberalism*. New York: Columbia University Press.
Raz J (1986) *The Morality of Freedom*. Oxford: Clarendon Press.
Rebonato R (2012) *Taking Liberties: A Critical Examination of Libertarian Paternalism*. Basingstoke: Palgrave Macmillan.
Saghai Y (2013) Salvaging the concept of nudge. *Journal of Medical Ethics* **39**(8), 487–493.
Schmidt A (2017) The power to nudge. *American Political Science Review* **111**(2), 404–417.
Sunstein C (1991) Preferences and politics. *Philosophy & Public Affairs* **20**(1), 3–34.
Sunstein C (2014) *Why Nudge? The Politics of Libertarian Paternalism*. New Haven, CT & London: Yale University Press.
Sunstein C (2015) Nudging and choice architecture: ethical considerations. *Yale Journal on Regulation* **32**, 413–450.
Sunstein C (2019) *On Freedom*. Princeton and Oxford: Princeton University Press.
Sunstein C and Thaler R (2003) Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review* **70**(4), 1159–1202.
Thaler R and Sunstein C (2008) *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.
Waldron J (2014) It's All for Your Own Good. *New York Review of Books*, 9 October.
Wall S (1998) *Liberalism, Perfectionism and Restraint*. Cambridge: Cambridge University Press.
Wilkinson TM (2013) Nudging and manipulation. *Political Studies* **61**(2), 341–355.
Williams B (1985) *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.
Willis L (2013) When nudges fail: slippery defaults. *The University of Chicago Law Review* **80**(3), 1155–1229.
Zajonc R (2006) Mere exposure: a gateway to the subliminal. In Lichtenstein S and Slovic P (eds), *The Construction of Preference*. Cambridge: Cambridge University Press, pp. 464–470.

Cite this article: Kniess J (2022). Libertarian Paternalism and the Problem of Preference Architecture. *British Journal of Political Science* **52**, 921–933. https://doi.org/10.1017/S0007123420000630