Behavioural Public Policy (2025), **9**, 249–269
doi:10.1017/bpp.2022.17

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Agentic preferences: a foundation for nudging when preferences are endogenous

Mark Fabian[1]* and Malte Dold[2]

[1]Bennett Institute for Public Policy, University of Cambridge, Cambridge, UK and [2]Department of Economics, Pomona College, Claremont, CA, USA
***Corresponding author:** Mark Fabian, email: mark.otto.fabian@gmail.com

(Received 10 October 2021; revised 18 May 2022; accepted 18 May 2022)

Abstract

Since the publication of the seminal book *Nudge* by Thaler and Sunstein, several critics have highlighted preference endogeneity as a serious obstacle to nudging. When individuals hold preferences that are dynamic and endogenous to the nudge frame, it is unclear what the normative benchmark for libertarian paternalistic policies should be. While acknowledging this issue, the pro-nudging camp has not yet sufficiently addressed it. This article aims to fill this void by presenting a conditional defence of nudging when preferences are endogenous. We explain the learning process through which individuals establish 'agentic' preferences: preferences that are sufficiently stable, reasonable, autonomous and associated with organismic well-being to ground the 'welfare' principle of libertarian paternalism. To describe this process, we draw on theories from psychological science, in particular self-discrepancy theory and self-determination theory. We argue that agentic preferences are not only welfare-relevant and thus appropriate to libertarian paternalism but can also be identified by choice architects.

Keywords: behavioural economics; welfare economics; endogenous preferences; self-determination theory; nudge; libertarian paternalism; preference formation

JEL Codes: D01; D91; I38

Introduction

Nudges are nowadays a ubiquitous feature of behavioural public policy. In the classic formulation of Thaler and Sunstein (2008), nudges are a small change to the 'choice architecture' facing an individual designed to counteract cognitive bias and thereby allow that individual to choose according to their 'true' preference in a given situation. Let's consider a common example, illustrated in Figures 1 through 3. Many people express an antecedent preference to eat healthy and consequently plan to consume vegetables (Figure 1). However, when passing by the confectionary in the office cafeteria on an empty stomach, they get a sudden craving for sugar and consequently consume a candy bar (Figure 2; an example of present bias). Placing the vegetables in

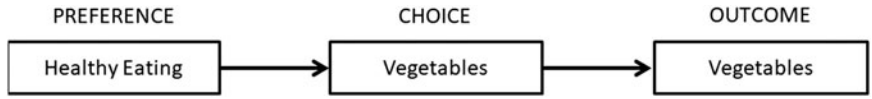

Figure 1. Scenario A – Classic revealed preference.

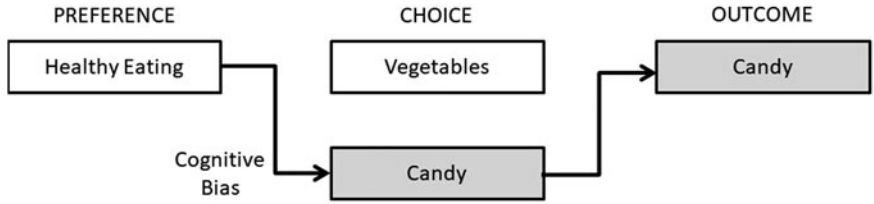

Figure 2. Scenario B – Revealed preference with cognitive bias.

an easy-to-access area of the cafeteria and the confectionary somewhere out of the way – a nudge – prevents this cognitive bias from triggering. The individual then purchases healthy food in line with their 'true' preference to eat healthy (Figure 3).

Nudges are normatively fraught because they involve paternalistic manipulation, which liberal societies tend to frown upon. Thaler and Sunstein (2008) developed the normative paradigm of 'libertarian paternalism' to identify arguably acceptable circumstances on the margin of paternalism where nudging is relatively less controversially. Libertarian paternalism has four principle components:

Cost-Effectiveness: Nudges should offer meaningful results at low cost, and their causal impact on behaviour should be well established using experimental evaluation methods like randomised control trials.

Autonomy: Nudges should preserve all available choices (as in the case of putting the candy out of the way). They should not be coercive or involve bans (that is a 'shove').

Transparent: Nudgers should be frank about their methods and objectives.

Welfare: Nudges can only be justified when they promote the preferences of the people being nudged.

Dold (2018), Sugden (2018), Rizzo and Whitman (2019) and among others have recently questioned the normative adequacy of the 'welfare' condition for justifying

Figure 3. Scenario C – Revealed preference with cognitive bias and nudge.

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

nudges because preferences are often dynamic and endogenous to the decision frame. This means that people's revealed preferences can be changed by the nudge itself. Without the redesign of the cafeteria, the individual appears to prefer confectionary. After the redesign, they choose to go without. Which preference is the individual's 'true' preference? When preferences are endogenous to nudges, the normative argument outlined above breaks down because there are no clear antecedent preferences.

Paul and Sunstein (2019) have recently argued that the welfare condition holds so long as nudged individuals assent to the nudge *post-hoc*. For example, upon learning that the cafeteria layout was altered to encourage healthy consumption, cafeteria patrons agree that this design decision was in their interest. This moves the 'welfare' principle away from 'true' preferences to the slightly different notion of 'better off *as judged by themselves*' (AJBT).

We find this proposal unsatisfying. It authorises a stronger form of paternalism – manipulatively altering an actor's behaviour on the grounds that it *will be* good for them. The AJBT principle opens the possibility of introducing nudges where no biases are present but are inferred to exist based on behaviour change associated with a nudge. In such cases, it is possible that the nudge *introduces* a bias. For example, Kallbekken and Sælen (2013) suggest using smaller plates in cafeterias to remove 'biased perceptions' arising from 'visual illusions' that lead to excessive portion size and consequent food waste. Coincidentally, they argue that customer satisfaction is unchanged by the plate size intervention – this is a form of justification by *assumed post-hoc* assent. Yet Sobal and Wansink (2007) argue that 'plate shape and size delineate norms for appropriate amounts of food to eat at a meal'. Using smaller plates might therefore simply be introducing social desirability bias into portion size decisions. Because neither antecedent preferences nor the existence of a bias is established, the plate size intervention may just be pure manipulation. This contravenes the 'libertarian' in libertarian paternalism, which requires not just the preservation of choice but also that the nudge helps the agent to realise their own preferences. If the agent has no such preferences, then the preferences of policymakers will often take their place. There may be ethical grounds to argue for such interventions, such as moral repugnance or market failures of information and externalities, but these will need to come from beyond the boundaries delineated by libertarian paternalism (Conly, 2012).

We explore a hopefully more fruitful approach to overcoming the issue of endogenous preferences. We explain how individuals can come to possess 'agentic' preferences that are sufficiently stable, reasonable, autonomous and associated with well-being to ground the 'welfare' principle of libertarian paternalism. We then discuss how these preferences can be identified by choice architects and folded into policy design.

Before explaining what agentic preferences are, we want to highlight what they are not: they are neither *total subjective comparative evaluations* (Hausman, 2012) nor *constructed preferences* (Bettman *et al.*, 1998). In the former case, preferences are seen as a result of a series of cognitively demanding operations on, among other things, the relevant partial evaluations (Angner, 2018). In the latter case, preferences are conceptualised as being assembled 'on the go' in various decision situations based on context cues; as a result, constructed preferences are typically context-dependent and unstable (Lichtenstein & Slovic, 2006). In contrast, agentic preferences can be the

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

result of experiential, intuitive learning and are not necessarily the outcome of a cognitively demanding process. They are also relatively stable across contexts and not prone to simple framing affects. The reason, as we will explain, is a distinct underlying process of preference formation.

Self-actualisation and agentic preferences

One channel through which individuals could develop preferences appropriate to nudges is what could be called 'education'. This involves engaging an agent's reasoning faculties in a process of conscious deliberation to arrive at preferences that they are confident are appropriate for themselves. For example, Thaler and Sunstein (2008, p. 111–112) refer to training individuals sometimes receive in planning for retirement that makes them aware of the need to contribute more to their pension funds. Thaler and Sunstein take such 'educated' or 'well-deliberated' preferences as justification for nudging people to save more using higher default savings rates.

The problem with education is that it often isn't compatible with nudging. Education is slow, conscious and heavily dependent on 'system 2' – frontal lobe processing. Nudges are quick and typical work with 'system 1' – instinctive processing (Kahneman, 2011). Indeed, education aligns more closely with the notion of 'boosting' (Hertwig & Grüne-Yanoff, 2017). Boosts actively engage the attention of citizens in order 'to hone the skills of the general public in dealing with risks and making decisions' (Hertwig & Grüne-Yanoff, 2017, p. 311). We think there is some potential for blending here. The identification of educated preferences could be used to inform choice architecture. But given the tension between education and nudging and the lively literature on boosting, we leave education out of our analysis here.

We focus instead on *learning*. By this, we mean a complex process of forming 'agentic preferences' through self-actualisation. To describe this process, we draw on a range of psychological theories, especially *self-discrepancy theory* and *self-determination theory*. We use the term 'agentic' preferences both to emphasise their connection to authenticity, agency and autonomy and to differentiate them from the related concept of 'well-laundered' preferences (Hausman, 2012). One can think of self-actualisation as a laundering process, but 'well-laundered' preferences are associated with approaches in behavioural welfare economics that rely on economic notions of rationality. The problem with these approaches is that they do not sufficiently explain why welfare-improving choices need to follow formal axioms of consistency (Dold, 2018). Our proposal, in contrast, follows Sugden's plea that behavioural welfare economists '[should] learn to live with the facts of human psychology' (2018, p. 82). By folding theories of well-being from psychology into our analysis of self-actualisation, we can explain the stability of agentic preferences and thus their suitability for the preference-satisfaction notion of welfare in nudging. We can also explain how agentic preferences relate directly to psychologically rich notions of well-being like positive mood, life satisfaction, vitality, the absence of psychopathology, and feelings of autonomy, competence, relatedness and purpose.

Where do preferences come from?

Fabian (2020) combines several theories of 'self' and motivation to develop a theory of self-actualisation in the context of well-being that he calls 'coalescence'. We adapt his

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

theory to provide some guidance on how agentic preferences form. We do not presume that this theory is complete, generalisable or easy to apply across areas of interest to economists. It does not provide much insight into why people choose one brand of butter over another, for example. However, it does provide insights into preference formation and change that are instructive in the context of behavioural welfare economics. In particular, it provides scaffolding upon which to build a psychologically realistic theory of endogenous preferences.

Self-discrepancy theory

The first layer of the theory comes from self-discrepancy theory (Higgins, 1987). Simplistically, self-actualisation is driven by attempts to align our 'actual' self with our 'ideal' and 'ought' selves. Our *actual self* is who we are right now, including certain innate motivations, talents and parameters like height. We may have an organic disposition towards science and adventure, for example. Our *ideal self* is who we would like to be, like an astronaut. The actual self encourages and constrains the ideal self. Someone predisposed to exercise will have an easier time becoming an athlete, for example. Equally, someone short will struggle to become a competitive basketballer. Finally, our *ought self* is who we feel we have a responsibility to be. For example, a religious individual might ideally like to complete several pilgrimages but feels that they ought to stay home and provide for their family. In reality, the self is fragmented and compartmentalised into many more 'multiple selves' than these three concepts (Showers & Zeigler-Hill, 2012), but the concepts nonetheless provide a useful and parsimonious framework through which to understand a central mechanism of preference formation. The ideal and ought selves have some similarities with what Callard (2018) calls as 'aspirational self'. They are valued identities that we only understand vaguely and reason towards proleptically, calibrating them as we learn about what is valuable and what would suit us through choice and associated feedback.

The ideal and ought selves are high-level goals of some sort, like becoming an astronaut. Various lower-level goals are then derived from this higher-level goal in a nested fashion. For example, to become an astronaut, one would need to be knowledgeable in physics and fit enough to withstand a space shuttle launch. To become fit, one would need to attend the gym regularly and eat well. A good diet would consist of vegetables and proteins while being low in fat and sugar. The individual would therefore prefer a chickpea salad over a burger and chips combo for lunch. In this way, the high-level goal informs *preferences* down to a minute level, including even seemingly trivial consumption choices, and the ideal self comes to constitute an entire identity. Note that neither the ideal nor ought self is necessarily ambitious. We use the example of an astronaut because it is straightforward to work with in our analysis, but many people have ideal and ought selves that are more mundane. Someone may ideally like to be a homebody with a well-kept garden, or a reliable father, or a good boss in small local firm.

The process of harmonising the actual, ideal, and ought selves through behavioural choices is guided by affective signals (Showers & Zeigler-Hill, 2012). Notably, discrepancies between the actual and ideal selves trigger depression, while discrepancies

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

between the actual and ought selves trigger anxiety (Silvia & Eddington, 2012). These psychopathologies are especially acute when the actual self aligns with the 'feared self', a conceptualisation of who the individual does not want to be (Woodman & Hemmings, 2008). Conversely, coinciding with your ideal self brings positive affect. We feel joy when we achieve our goals (Emmons, 1986). We feel exhilarated when we recognise that we are progressing dynamically towards those goals. And we feel vitality or a sense of easy motivation when we are working towards valued ends (Sheldon & Elliot, 1999). Note that affective signals come when choices are realised, which is often a process rather than an event. A burger might taste delicious when ingested, but some of the affective signal associated with consuming the burger comes later when you're looking at your gut and feeling sad about your lack of progress towards your aesthetic goal. Learning and preference formation thus encompasses much more than decision utility.

Self-actualisation through harmonisation of the self-constructs proceeds by way of choice, feedback and adjustment (illustrated in Figure 4). The individual begins by developing some conception of the ideal self that is amenable to the ought self and not ruled out by parameters of the actual self. They then try to become this ideal self by choosing in accordance with the values, behaviours and symbols inherent to that ideal self. In other words, they try to affirm an identity. For example, the wannabe astronaut might take steps in that direction by joining a gymnastics club and taking a course in astrochemistry. Their choice behaviour over time will reveal their actual self and make apparent whether they are progressing dynamically towards their ideal self. Imagine that the wannabe astronaut always finds excuses to skip out on gym training but loves their astrochemistry course. They will be revealed in their actions as not being their ideal self, namely on-track to be an astronaut. The individual must introspect on the affective signals that will accompany this information. If they become depressed, as predicted by self-discrepancy theory, it suggests that this ideal self-concept (astronaut) might be appropriate for them, but they must work

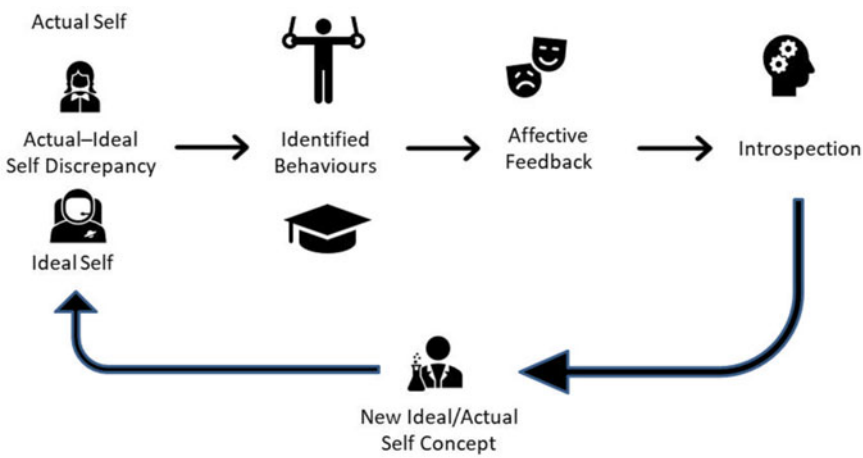

Figure 4. The iterative process of self-actualisation.

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

harder or smarter to realise it. Perhaps they need to swap from gymnastics to swimming. If they are not upset, then perhaps they have little motivation for being an astronaut and should consider an alternate ideal self that is more inspiring. Given the positive feedback they get from astrochemistry, maybe a planet-bound research role will suit them better. They should pursue these feelings further by deepening the associated identity, perhaps by seeking employment as a research assistant, or taking similar courses.

Self-determination theory

The second layer of our theory of preference formation comes from self-determination theory (Ryan & Deci, 2017; SDT), which helps to describe the nature of the actual, ideal and ought self-constructs in terms of motivational differences. It also provides some explanation as to why some values and behaviours 'fit' an individual better than others.

SDT posits a spectrum of motivation running from intrinsic on one end to extrinsic on the other (see Figure 5). Intrinsically motivated behaviours are engaged in for their own sake, often spontaneously. They are thus self-*determined*. In contrast, extrinsically motivated behaviours require some degree of self-*regulation* because resistant parts of the psyche must be suppressed to undertake such behaviours. The most extreme form of extrinsic motivation is *duress*, where an outside influence coerces behaviour – these are *controlled* behaviours and are not engaged in autonomously. One step closer to intrinsic motivation is *introjected* behaviours. These are extrinsically motivated, but unlike duress, here the individual administers the control upon themselves. An illustrative example is behaviours done solely to earn parental approval. *Identified* behaviours are recognised as valuable by the individual, but they are not intrinsically motivated. For example, many people exercise to improve their health, but have only limited motivation for the exercise itself. Extrinsic, introjected and identified behaviour all involve degrees of self-regulation because they are not undertaken spontaneously as a result of intrinsic motivation. *Integrated* behaviours have been connected to other intrinsically motivated behaviours. They thereby become easier to motivate and move into a more central position within an individual's identity. For example, someone who joins a soccer club for exercise may find that 'doing soccer' becomes connected over time to socialising, watching soccer and other behaviours and values that are intrinsically motivated for them. This makes 'doing soccer' easier to motivate.

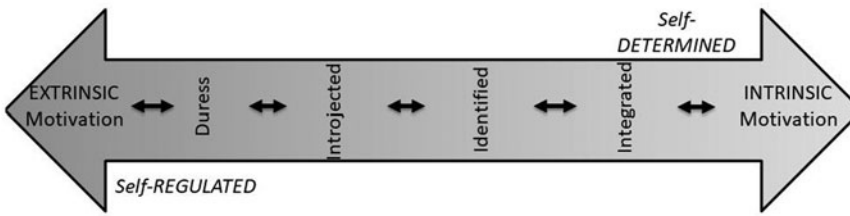

Figure 5. The motivation spectrum in self-determination theory.

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

The actual, ideal and ought selves map neatly but imperfectly onto SDT's motivation spectrum (see Figure 6). The ought self involves regulating one's behaviour to conform to ethical objectives, including other-regarding preferences. It thus corresponds to introjected and identified motivation. The ideal self is something you want to become. It is made up of identified behaviours that become integrated over time. For example, an aspiring astronaut might identify with fitness and physics literacy. Pursuing these behaviours is often arduous and requires self-regulation (i.e., willpower), especially in the early stages (Besser-Jones, 2014). The individual must master difficult mathematics and train up to an athletic standard well beyond their current capacity. However, as the individual's competence improves and these behaviours become assimilated into the individual's life and routines, they become *integrated*. The individual might incorporate cycling into their commute to the lab, for example, and see the world in a more enchanting way thanks to their newfound mathematical literacy. This makes their motivation easier. Finally, once the individual becomes their ideal self, the associated behaviours may become fully *intrinsic*. Once they are a skilful physicist and apex athlete, the associated tasks will no longer seem arduous. Indeed, the individual will likely get a rush of positive affect when executing these tasks because of feelings of competence. We can see here that varieties of motivation are themselves a kind of affective signal.

SDT argues that the objective of motivation is the nourishment of basic psychological needs for *autonomy*, *competence* and *relatedness*. Autonomy is the sense of being volitional in one's life. Competence is the sense that one is skilful at behaviours that promote one's flourishing. And relatedness is the sense of being loved and cared for, especially by valued others, and of belonging. Experimental studies across multiple national and cultural contexts have evidenced the claim that nourishing basic psychological needs increases subjective well-being (SWB: a combination of positive affect and life satisfaction), vitality and self-esteem and ensures ease of motivation, as well as reducing depression, anxiety, compartmentalisation, defensiveness and personal rigidity (Sheldon *et al.*, 2004, 2009; Church *et al.*, 2013; Chen *et al.*, 2015). These indicators of need satisfaction, many of which are feelings, are analogous to *experienced utility*.

Introjected and identified behaviours can yield situations where the *expected* decision utility associated with a particular behaviour was larger than the *actual*

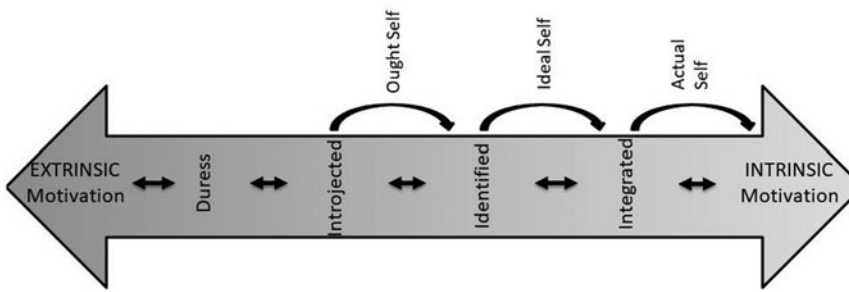

Figure 6. Self-concepts on the spectrum of motivation.

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

experienced utility (Kahneman *et al*., 1997). We might expect that dressing fashionably will have high utility because it will make us popular (an introjected motivation), but this turns out to be incorrect because our new friends are as shallow as the identity we express in our clothes. Similarly, we might think that classical music is enriching (an identified motivation) but we find it boring in practice.

Such discrepancies between expected decision utility and actual experienced utility in the SDT framework and their relationship to motivation, behaviour and well-being are a stark theme of studies of intrinsic and extrinsic *pursuits*. Empirically, goals associated with extrinsic and contingent rewards like money, fame and image provide relatively weaker payoffs than goals associated with intrinsic rewards like enjoyable experiences, personal growth and intimacy (Kasser & Ryan, 1993, 1996). This is because extrinsic pursuits do not nurture basic psychological needs (or, when they do so, then only to a smaller degree).[1] Sheldon has expanded these insights into a more general theory of *goal self-concordance* (Sheldon & Elliot, 1999). He finds that people derive greater psychological well-being from goals that fit their personalities, or more daringly, their innate selves (Sheldon & Vansteenkiste, 2005). The 'innate self' implied here is very minimalist. It corresponds only to some elementary predispositions, aptitudes and biological parameters like height and intelligence. It is a close conceptual analogue of the 'actual self' at the onset of adulthood before self-actualisation starts to change the actual self as it harmonises with the ideal and ought selves. The innate self of SDT certainly has little relationship to the notion of 'true' preferences in behavioural welfare economics (Dold, 2018).

Social feedback

Alongside affective feedback, agents can also receive social feedback from choice and behaviour. Indeed, the two varieties of feedback can be bound up together. Guilt, shame and self-esteem are often described as 'social emotions' that communicate to us when we are doing things that will be negatively perceived by our peers (Leary & Baumeister, 2000; Haidt, 2012; Leary, 2012). Strictly social feedback can come in a variety of forms including dis/approval from our peers, encouragement and advice. Social appraisals have been found to exert more influence on self-appraisal when the perceiver is considered by the perceived to be relevant to their self-concept, an in-group member, desirable, valued or otherwise important (Alicke *et al*., 2012; Wallace & Tice, 2012). Neuroscience studies align with this result. They show that when an appraiser is from a group you care about, their appraisals of you will activate the self-assessment part of your brain. This is not the case when they are from a group you don't care about or a random stranger (Devos *et al*., 2012, p. 158).

One of body of literature in the 'social self' space within psychological science that is especially relevant to self-actualisation is 'self-verification' theory (Swann, 2011). It conjectures that people seek confirmation of their self-views from others. One way we come to know whether we have aligned with our ideal and ought selves is if our peers

[1]Evidence suggests that this poor relationship between extrinsic aspirations and SWB holds even in social contexts like business schools and corporate law firms that espouse extrinsic aspirations like money and power and celebrate their achievement (Kasser & Ahuvia 2002; Vansteenkiste *et al*., 2006).

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

avow that we have. Empirical studies in self-verification theory find that self-views guide social interaction and, provided they are stable, make an individual's behaviour more predictable to others. This predictability stabilises the way others respond to the individual, which makes it easier to verify one's self-view through social interaction. Stable self-views thus encourage the emergence of a stable, coherent, social environment and vice versa, leading to a virtuous cycle wherein both self-concept and social environment become clearer and better fitted to each other. An important, empirically validated hypothesis that emerges from this theory is that people prefer social appraisals that align with their self-view even when these appraisals are negative (Swann & Buhrmester, 2012). People move away from both incorrect and correct-but-negative appraisals over time towards groups that are both affirming and accurate in their social appraisals.

Social feedback is especially relevant to the ought self. Guilt and shame are morally coded emotions, and social approval of conduct is a central theme in many accounts of ethical behaviour, such as Adam Smith's *Theory of Moral Sentiments*. The introjected 'responsibilities' that make up much of the ought self are typically tied to social expectations. We feel a duty to conform to certain ethical precepts, and deviation from these precepts can result in anxiety owing to an actual–ought self-discrepancy. This is one way in which other-regarding preferences enter into self-actualisation, but such preferences need not be introjected (Besser-Jones, 2014). Individuals can identify with certain ethical values and other-regarding preferences and pursue these authentically and autonomously as part of self-actualisation. This appears to be the case with converts to many ethical systems, both religious and secular, such as the growing membership of the effective altruism movement. Such individuals desire to be 'good people'. Over time, such identified values can even become intrinsically motivated through the process of internalisation. Indeed, some people may even possess intrinsically motivated other-regarding preferences as part of their innate selves. Experiments in game theory, for example, have identified 'supercooperators' who seem to easily engage in generous behaviour (Novak & Highfield, 2011).

Agentic preferences

The idea that affective signals guide behaviour and shape preferences over time is central to both self-discrepancy theory and SDT. In the former, discrepancies between self-concepts give rise to negative affect, while coincidence between the actual and ideal selves is punctuated by positive affect. In SDT, intrinsically motivated activities are associated with positive emotions and the absence of psychopathology, in contrast to extrinsically motivated activities. These affective signals encourage the individual to detach from extrinsic pursuits and unsuitable conceptualisations of the ideal self over time and comport towards identified and especially intrinsically motivated activities that characterise a self-congruent ideal self. This is how *agentic preferences* form and come to be understood and acted on by an individual.

We describe the preferences that emerge from the self-actualisation process outlined above as 'agentic' to emphasise their relationship to valuation, self-expression and autonomy. We take our notion of agency from Sen (1985, pp. 203–204), who

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

defines it as 'the freedom to achieve whatever the person, as a responsible agent, decides he or she should achieve'; he continues in saying that '[a] person's agency aspect cannot be understood without taking note of his or her aims, objectives, allegiances, obligations, and – in a broad sense – the person's conception of the good'. Our analysis speaks to Sen's themes. Autonomy and the pursuit of intrinsic motivations and identified values speak to 'freedom to achieve whatever the person … decides he or she should achieve'. The introspection and calibration that is fundamental to our notion of self-actualisation implies reflective endorsement of preferences, the proffering of reasons for those preferences and taking 'responsibility' for choices on the basis of those preferences. 'Aims and objectives' are preferences, 'allegiances' come in through the need for relatedness, and 'obligations' are present in the ought self. Harmony between the actual, ideal and ought selves requires alignment between the person's behaviour and their 'conception of the good'. Note that we are not making any claims about what is required for 'agency' here. We are merely justifying our use of the term in a descriptive capacity.

The welfare-relevance of agentic preferences

The relevance of agentic preferences to welfare can be justified through multiple channels (Fabian, 2020). To start, self-actualisation nourishes basic psychological needs. This is a prominent account of well-being in psychology that has been linked to high life satisfaction (Martela & Sheldon, 2019), another account of well-being common to both psychology (Diener *et al.*, 2009) and philosophy (Sumner, 1996). Self-determination theory presents itself as a *eudaimonic* theory of well-being, and Besser-Jones (2014) has explicated the extent of its compatibility with Aristotelian notions of well-being. Self-actualisation in our model is guided by affective signals. As it matures, individuals will detach from values and behaviours associated with negative affect and deepen their engagement with those associated with positive feedback. In this way, self-actualisation leads to what Kahneman (1999) calls 'experienced utility' or 'objective happiness', and satisfies most hedonistic accounts of well-being (Feldman, 2002). One does not need to endorse any of these accounts of well-being to utilise agentic preferences as a foundation for the welfare principle of nudging. However, it is encouraging that our account is compatible with so many perspectives on well-being, and with psychologically realistic accounts of human motivation.

Of course, the account of well-being most relevant to economics and nudging is that well-being consists in the satisfaction of preferences, and our claim is that agentic preferences are the relevant kind of preference for justifying a variety of nudges. Desroches (2020) has previously argued that 'value-based preferences' like those articulated by Tiberius (2018) provide a 'rigorous way for thinking about classic choice situations that have long interested behavioural economists and philosophers of economics'. She argues (p. 13) that 'well-being consists in the fulfilment of an appropriate set of values over a lifetime'. In turn, 'appropriate values are (1) suited to our desires and emotions, (2) reflectively endorsed, and (3) capable of being fulfilled together over time … appropriate values are objects of relatively sustained and integrated emotions, desires, and judgements' (p. 41). Our agentic preferences are

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

highly compatible with Tiberius' theory[2] and are psychologically sophisticated kind of 'value-based preferences'. Succinctly, as the self-actualisation process is guided by affective signals, the resultant agentic preferences are 'suited to our desires and emotions'. The introspective element of self-actualisation implies 'reflective endorsement'. And the need to harmonise the actual, ideal and ought selves implies that agentic preferences can be 'fulfilled together over time'. Incompatible self-concepts will lead to compartmentalisation and dissonance, which does not nourish needs and impinges on integration and self-actualisation. Our emphasis on *process* and *learning* and the notion that agentic preferences are refined and harmonised over time explains why agentic preferences are 'objects of sustained and integrated emotions, desires and judgements'.

Our model of agentic preferences can be more thoroughly linked to the notion of preferences utilised in economic decision theory. The process of self-actualisation through goal approach guided by affective signals can be simplistically represented as follows:

1. The individual affirms a desired identity 'or self' (a preference) in choice.
2. Social and affective feedback is received as a consequence of that choice. The nature of this feedback will depend on whether the choice harmonises the actual, ideal and ought selves, whether it nourishes basic needs, and whether it is self-congruent.
3. The individual introspects on this feedback and adjusts their preferences accordingly.
    a. Positive feedback promotes further investment in the identity ('doubling down').
    b. Negative feedback promotes disengagement from or reform of the identity.
    c. Feedback as expected promotes maintenance of the status quo.
4. Repeat from 1, as depicted in Figure 4.

This model of preference learning can be reframed in terms of expected decision utility and actual experienced utility. Identities are collections of nested goals and pursuits, which give rise to preferences associated with *expected* decision utility. Achievement of goals is then synonymous with preference-satisfaction, which provides *actual* experienced utility (though this utility is not instantaneous). If experienced utility is greater than expected utility, then the individual will likely double down on that activity. For example, if someone striving to be an astronaut finds that they enjoy athletics and physics, they may not only do more of these activities but also take on other behaviours associated with being an astronaut, like scuba diving to become familiar with different environmental pressures. In contrast, if experienced utility is less than expected utility, then individuals will likely disengage from the behaviour over time. If the wannabe astronaut finds mathematics boring and arduous, they will give up on being an astronaut.

[2]Indeed, Fabian (2022) extensively integrates his (2020) model of self-actualisation, which we use, with Tiberius' theory of well-being as value-fulfilment.

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

When agentic preferences are satisfied, there will be limited incongruence between expected utility and experienced utility, and associated behaviours will be intrinsically motivated. This is because such preferences have gone through the 4-step learning process above repeatedly. Individuals disengage from preferences with negative feedback, and further invest in activities associated with positive feedback until they start to get negative feedback associated with overinvestment. The wannabe astronaut would not get so carried away with athletics that it interferes with their physics studies, for example. This implies that agentic preferences and the routine behaviours through which they are satisfied are largely *stable*.

When experienced utility repeatedly approximates expected utility, the associated behaviours gradually become 'automatised' (Bargh & Chartrand, 1999; Kruglanski & Szumowska, 2020). What this means is that they are done instinctively – what Kahneman (2011) would call system 1 processing – rather than consciously. Introspection is not required, and so system 2 (conscious processing) is unnecessary. However, if the individual were to reflect on their automatised behaviours, 'they would fit with one's values or needs and could readily be changed when they no longer fit' (Ryan & Deci 2004, p. 448). Agentic preferences then are automatic, stable, based on repeated rounds of choice and information gathering (i.e., they are 'informed'), reasoned, self-congruent, intrinsically motivated and nourish basic psychological needs. They thus satisfy almost all the conditions typically associated with 'welfare-relevant' preferences according to economic philosophers. The only exception is that they do not necessarily conform to the highly stylised and mathematised notions of economic 'rationality' that stress things like transitivity and completeness. These can instead be taken *as assumptions in modelling*, thereby allowing agentic preferences to be plugged into much of the microeconomic architecture developed out of the notion that 'utility' is 'preference satisfaction'.

The theory of preference formation we have outlined provides criteria upon which to assess whether preferences are agentic and thus suitable as a welfare criterion. First, agentic preferences have been through a *learning process* of choice, (affective) feedback and introspection. This process is evident in negligible choices like how we take our coffee through to more significant issues like occupational choice. Second, agentic preferences are generally *nested within higher-level goals*. We earlier gave the example of food consumption choices being nested within athletic goals, which were in turn nested within career ambition. Third, agentic preferences connect to form a *coherent identity*. The identity of 'athlete', for example, informs choices over food, clothing, leisure and occupation. Compartmentalisation, contradictory values and inconsistencies between values and behaviours imply incomplete self-actualisation and thus relatively less agentic preferences (they may still be sufficient to inform nudges, but the exploration of such nuances is beyond the scope of this article). Fourth, *affective feedback* from the satisfaction of agentic preferences will be largely in line with expectations. Surprise, regret, disappointment and the like thus imply incomplete self-actualisation and not (yet) agentic preferences. The satisfaction of agentic preferences will almost invariably be associated with positive affective feedback overall and in the long run. Fifth, agentic preferences tend to manifest as *repeated choices* because people detach over time from choices that provide negative affective feedback while comporting towards choices that provide positive affective

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

feedback. Sixth, agentic preferences tend to manifest as *automatised choices*. Prior experience has taught the agent that the choice delivers positive payoffs in line with expectations, and so they do not reconsider their choice unless unexpected payoffs materialise. In this sense, agentic preferences are 'stable'. Uncertainty, trepidation and extensive prior research imply the absence of agentic preferences. And, finally, seventh, people will be *more easily motivated* towards agentic preferences because they are associated with relatively intrinsic forms of motivation like identification and integration. Reluctance, the engagement of willpower, amotivation and the like imply incompletely agentic preferences.

We want to stress that these criteria do not all need to be evidently met for a choice architect to be satisfied that they have identified agentic preferences. The realities of policymaking mean that identifying such a variety of largely internal psychological features of agents may be practically very challenging. We list them instead as a collection of indicators that choice architects can draw on, within the context of their potential nudge, to consider whether agentic preferences are present. If very few of these criteria are met, then it should discourage policymakers from relying on libertarian paternalism as a source of justification. When school-leavers choose an undergraduate degree, for example, they have had few if any opportunities for repeated choice and associated feedback, their identity is nascent, and their choice is typically highly considered rather than automatised. It is unlikely that they are acting to satisfy an agentic preference. In contrast, when they are considering what music to listen to, a relatively agentic preference is much more likely.

A related concern for identifying agentic preferences is apparent variability in choices across contexts. For example, well-known YouTube fitness influencer Jeff Cavalier maintains a strict diet grounded in his agentic preferences for health, fitness and lean aesthetics, but eats one slice of carrot cake per annum. A choice architect observing this might perceive it as a deviation from Cavalier's agentic preferences and assume a cognitive bias is interfering with his welfare. This seems unlikely to be the case. Cavalier's decision occurs during family Thanksgiving. Perhaps social desirability bias is at play. But given how considered the decision is, it seems more reasonable to conclude that Cavalier has agentic preferences for *both* participating in Thanksgiving the way his family prefers (ought self), and maintaining his lean physique (ideal self). He has harmonised these two agentic preferences by associating them with particular contexts, and by maintaining a consistent level of variation in his choice behaviour. There is nothing inherently unstable or irrational about satisfying different agentic preferences depending on the context. Furthermore, one can have an agentic preference for variety and experimentation. Our psychologically rich account of how agentic preference form through self-actualisation reveals that the one-shot scenarios that are the bread and butter of behavioural economics are potentially misleading with regard to the stability of people's preferences. Preferences can be varied in stable ways and context-dependent in structurally rational (Broome, 2013) ways without compromising fundamental assumptions of rational choice theory.

In fact, the psychological literature on 'multiple selves' empirically validates this claim (Showers & Zeigler-Hill, 2012). From adolescence onwards, individuals develop 'a dramatic rise in the detection of contradictory self-attributes that lead to conflict

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

and confusion' (Harter, 2012). The most common way of integrating these multiple selves is to determine which self-concept is most appropriate for what context. Showers and Zeigler-Hill (2012) offer the example of a 'superdad' who is a nurturing father at home but a hard-arsed executive at the office. The superdad will determine, through rational introspection on affective and social signals, when to engage his nurturing or cutthroat personas or his more cutthroat persona. His children might not take kindly to an executive style of household management, for example. Their distress would communicate to him that he is failing to fulfil his superdad value and provoke behaviour change. His behavioural inconsistency across contexts actually reflects complex but stable and rational agentic preferences. This seeming instability might make identifying agentic preference and designing nudges on their basis more challenging, but we do not see this as a problem. Furthermore, in contexts where an individual holds conflicting or incompatible agentic preferences, we think, *in the first instance,* that nudgers should stay back and wait for the individual themselves to sort out this incompatibility to their satisfaction. There is too much ambiguity for the nudger with respect to their ability to confidently identify a stable preference (Read, 2006). Policymakers should always exercise caution when acting paternalistically, libertarian or otherwise.

Back to nudges

We have established how self-actualisation can lead to the emergence of relatively stable preferences that are sufficiently related to welfare to form the basis for nudging. We have also enumerated criteria that can be used to identify such 'agentic' preferences. Incorporating this learning process (as well as education, which we mentioned briefly) into our introductory example of vegetables versus candy yields (Figure 7). The central insight of Figure 7 is that in order to identify a welfare-relevant preference that can act as a foundation for nudging we need to model how that preference forms; we cannot simply take it as exogenous.

In the context of nudging, the agentic preference depicted in Figure 7 can effectively function as a welfare criterion, but it cannot be accessed through the traditional approach of revealed preference. This is because choice, and thus *revealed* preference, remains endogenous to the presence of biases and nudges. This endogeneity cannot be removed. However, what is relevant to welfare is that the underlying agentic

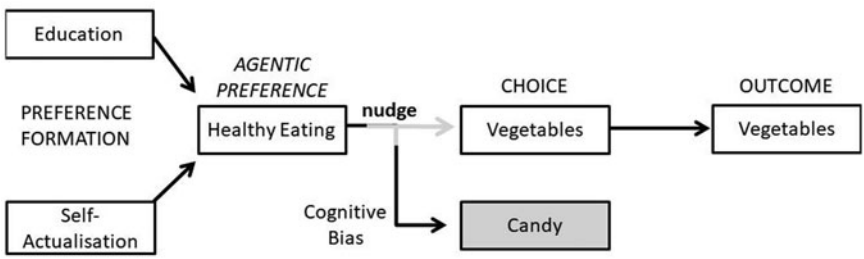

Figure 7. Revealed preferences with nudge, cognitive bias and preference laundering.

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

preference is not endogenous to nudging, only the revealed preference is. The 'welfare' principle of libertarian paternalism can therefore be satisfied so long as the presence of agentic preferences and cognitive biases is established *prior* to the introduction of a nudge. The immediate challenge is how to identify whether agentic preference satisfaction is being thwarted by cognitive biases if the revealed preference is ruled out as a methodology. We can't just ask people: one of the principal reasons to focus on revealed preference in welfare economics is because people's stated preferences often diverge from those observed in actual choices. We discuss some alternate strategies for nudging in the context of agentic preferences below.

Strategies for working with agentic preferences

First, policymakers could observe experienced utility in the form of affective feedback from choices made with and without nudges wherein cognitive biases are suspected. If a cognitive bias causes people to choose out of line with their agentic preferences, then they will feel *bad* after choosing. This is distinct from them feeling *good* or ambivalent after being nudged, as suggested by Paul and Sunstein (2019) in their 'as judged by themselves' reformulation of the welfare principle. Note, however, that negative feedback is only a necessary condition for introducing a nudge because it implies that the 'wrong' preferences were satisfied by a choice, but it is not sufficient. It suggests that a bias might be interfering with the satisfaction of an agentic preference, but it does not establish the existence of said bias.

Proving the existence of an interfering bias and the effectiveness of a nudge would require experimental conditions of choice without bias, choice with bias and choice with bias and nudge. In the absence of bias or nudge, individuals should choose their preference. The introduction of the cognitive bias, say by placing confectionary near the check-out, should then interfere with this preference satisfaction. Participants will report negative feedback from their choices. Implementing the nudge should then return choice to where it was without the bias or the nudge. Respondents will report no meaningful feedback because their expected utility from choosing in line with their preference aligns with their experienced utility. This approach is similar to how behavioural experiments in the nudge tradition have proceeded, except that agentic preferences are used here rather a neoclassical notion of 'rational' preferences. This is messier but more realistic – it takes psychology seriously.

A second strategy is *pre*-hoc assent. If you can demonstrate to people that they will likely fail to satisfy their agentic preference at the moment of choice due to a cognitive bias, many will assent to a nudge being implemented. For example, imagine an organisation like a gym whose members have agentic preferences for exercising. Now imagine that they are then shown compelling evidence from the cognitive biases literature that they are more likely to take the stairs if those stairs are painted like piano keys and make sounds when stepped on. We might reasonably expect the members to recognise this as a good outcome from the point of view of their agentic preferences for exercising and assent to the stairs being redesigned. This is a similar proposal to those involving strong 'transparency' within the libertarian paternalist paradigm. This principle requires policymakers to inform people about the nature and mechanism of

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

a nudge before implementing it. In an experimental study, Paunov *et al.* (2019) found that more stringent transparency can actually increase the uptake of nudges.

An important issue for policy applications of agentic preferences is that they may be robust to some forms of cognitive bias, but still susceptible to others. This can inform which nudges they are relevant to. For example, an individual committed to their agentic preferences may have the integrity to hold to them despite peer pressure and is thus less likely resistant to social desirability bias. However, this same conviction may make the individual more susceptible to confirmation bias and motivated reasoning when a desire to protect one's agentic preferences or self-image can distort the recruitment and evaluation of evidence (Nickerson, 1998; Epley & Gilovich, 2016). Additionally, the organismic drive to self-actualise (Ryan & Deci, 2017) has also evolved in the same context that produced the various biases of statistical reasoning catalogued by behavioural economists, such as conjunction fallacy, base rate fallacy, authority bias and framing effects (Kahneman, 2011). Finally, self-actualisation is an iterative process that takes place over potentially medium to long-time horizons, which makes agentic preferences acutely sensitive to biases relating to poor prospecting, such as default bias and hyperbolic discounting. An important inference from these observations is that agentic preferences do not exhaust the grounds on which a nudge could be justified. Nudges to overcome default bias in pension plans, for example, are more straightforwardly justified by appeal to a traditional 'rational' account of preferences (Hausman, 2012).

Considerations for policy practice

Our conditional defence of nudging raises at least two practical issues for nudges in policy practice related to cost. Recall that one source of legitimacy for nudges is their cost-effectiveness. They must be demonstrated to 'work' through experimental empirical designs, and they must be cheap to implement. Our proposed logic undermines this cost-effectiveness in two ways. First, it requires policymakers to exert far more effort than they presently do to establish the existence of agentic preferences and some cognitive bias that is interfering with the satisfaction of those preferences. Second, it requires policymakers to conduct more surveys than they currently do, either to establish negative feedback as evidence of some interference from cognitive bias, or to establish pre-hoc assent to nudges going forward.

We do not see the first issue as especially problematic; quite the opposite. It is unethical to introduce nudges, at least under the auspices of libertarian paternalism, without first checking for agentic preferences and the existence of some cognitive bias. Doing so runs the risk of simply nudging behaviour to align with the preferences of policymakers, perhaps even by *creating* a bias rather than removing one. For example, as part of a partnership with local councils in the UK, environmental organisation Hubbub installed 'ballot bins' that allow smokers to vote in spot polls using their cigarette butts. This nudges them to bin their butts rather than littering them. A good outcome for the council. But what preference do the smokers have? To litter owing to convenience, presumably. And what cognitive bias is being addressed? It seems that Hubbub *introduces* a cognitive bias around competitiveness to steer the behaviour of smokers in the direction Hubbub prefers. We have chosen this example

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

deliberately because the outcome seems so 'good'. Smokers amuse themselves, councils have less cleaning to do, and streets are cleaner. Yet this policy seems to *deliberately alter preferences* and thus contravenes the welfare principle of libertarian paternalism. If a psychological intervention leverages cognitive bias rather than removing them then it is not a nudge, just paternalism, or even pure manipulation. It may be possible to justify such policies, but not under the auspices of libertarian paternalism; it would rather be a political justification of nudging (Guala & Mittone, 2015). Notably, litter is a classic case of a negative externality where rational behaviour on the part of the litterers leads to net social costs. Hicks-Kaldor efficiency, among other normative principles, can be used to justify interventions against litter, including manipulative ones like ballot bins.

The additional costs associated with surveying can be minimised through effective sampling. If a representative sample (at least enough to provide statistical power) of the target population assents to the introduction of a nudge, then it can be rolled out to the population at large. For example, imagine a large university conducts a trial involving a representative sample of student volunteers that establishes agentic preferences for moderating caloric intake among a supermajority of those students. The study further establishes that cognitive biases emerging out of large plates in the cafeteria are interfering with the satisfaction of these agentic preferences. The university could now publicise the results of the trial to the entire student population and ask them, as part of a poll, whether they would like cafeteria designs nudged to make healthy eating choices easier. While far from free, these are low-cost methods for establishing legitimacy. In some cases, it may be straightforward to survey the entire population, as in the case of small and medium enterprises considering the implementation of nudges for their staff. Such spending seems appropriate when choice architects are trying to justify paternalistic manipulation. Indeed, Schmidt (2017) has previously argued that such quasi-democratic accountability upon nudgers is appropriate and could actually legitimise the more widespread use of nudges. We make no comment here on what sample size and representativeness in a survey/vote would be required to establish assent, though a simple majority strikes us as too low a bar to implement libertarian paternalistic interventions.

Conclusion

We have explicated a foundation for the 'welfare' principle of libertarian paternalism that overcomes the problem of endogenous preferences. In doing so, we have developed a model of preference formation that should be of interest to economists more broadly and welfare economists especially. Our model of preference formation is psychologically realistic and relevant to multiple accounts of well-being across economics, psychology and philosophy. Speaking simplistically, 'agentic preferences' form through an iterative and proleptic process of identification, choice, feedback, introspection and refinement. This process continues until the expected/decision utility associated with preference satisfaction aligns with the actual experienced utility of satisfying those preferences. At this point, the preference and associated behaviours become automatised. This process ensures that preferences integrate emotions, motivations and cognitions. Agentic preferences are stable, well-informed, related to well-

https://doi.org/10.1017/bpp.2022.17 Published online by Cambridge University Press

being, reasoned and individualised. They thus meet all the classic criteria for welfare-relevant preferences discussed in economic philosophy other than 'perfect' information and 'perfect' rationality. We think these agentic preferences could form a useful foundation for behavioural welfare economics going forward. However, we recognise that there is substantial work to be done applying our largely theoretical analysis to the practical realities of policymaking. In particular, our model is very individualistic, but nudges in practice tend to be targeted at larger entities like cafeteria patrons, gym members or college students. Yet, given self-selection effects, heterogeneity in agentic preferences might not be as big of a problem in those settings. Furthermore, since individuals are exposed to the same choice environment, they might in fact share similar biases. Admittedly, these are theoretical conjectures. We hope to address these issues in greater detail in future work and welcome robust debate in the meantime.

Competing interest. The authors have no competing interests to declare.

References

Alicke, M., C. Guenther and E. Zell (2012), 'Social Self-Analysis: Constructing and Maintaining Personal Identity', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York: Guildford, 291–308.

Angner, E. (2018), 'What preferences really are', *Philosophy of Science*, **85**(4): 660–681.

Bargh, J. and T. Chartrand (1999), 'The unbearable automaticity of being', *American Psychologist*, **54**(7): 462–479.

Besser-Jones, L. (2014), *Eudaimonic Ethics: The Philosophy and Psychology of Living Well*. London: Routledge.

Bettman, J. R., M. F. Luce and J. W. Payne (1998), 'Constructive consumer choice processes', *Journal of Consumer Research*, **25**(3): 187–217.

Broome, J. (2013), *Rationality Through Reasoning*. Hoboken: John Wiley & Sons.

Callard, A. (2018), *Aspiration: The Agency of Becoming*. New York City: Oxford University Press.

Chen, B., M. Vansteenkiste, W. Beyers, L. Boone, E. Deci, J. Van der Kaap-Deeder, B. Duriez, W. Lens, L. Matos, A. Mouratidis, R. Ryan, K. Sheldon, B. Shoenens, S. Van Petergem and J. Verstuyf (2015), 'Basic psychological need satisfaction, need frustration, and need strength across four cultures', *Motivation and Emotion*, **39**(2): 216–236.

Church, A., M. Katigbak, K. Locke, H. Zhang, J. Shen, J. de Jesus Vargas-Flores, J. Ibáñez-Reyes, J. Tanaka-Matsumi, G. Curtis, H. Cabrera, K. Mastor, J. Alvarez, F. Ortiz, Y. Simon and C. Ching (2013), 'Need satisfaction and well-being: testing self-determination theory in eight cultures', *Journal of Cross-Cultural Psychology*, **44**(4): 507–534.

Conly, S. (2012) *Against Autonomy: Justifying Coercive Paternalism*. Cambridge: Cambridge University Press.

Desroches, T. (2020), 'Value commitment, resolute choice, and the normative foundations of behavioural welfare economics', *Journal of Applied Philosophy*, **37**(4): 562–577. doi:10.1111/japp.12418.

Devos, T., Q. Hyunh and M. Banaji (2012), 'Implicit Self and Identity', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York City: Guildford, 155–179.

Diener, E., R. Lucas, W. Schimmack and R. Helliwell (2009), *Well-Being for Public Policy*. Oxford: Oxford University Press.

Dold, M. F. (2018), 'Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics', *Journal of Economic Methodology*, **25**(2): 160–17.

Emmons, R. (1986), 'Personal strivings: an approach to personality and subjective well-being', *Journal of Personality and Social Psychology*, **51**(5): 1058–1068.

Epley, N. and T. Gilovich (2016), 'The mechanics of motivated reasoning', *Journal of Economic Perspectives*, **30**(3): 133–40.

Fabian, M. (2020), 'The coalescence of being: a model of the self-actualisation process', *Journal of Happiness Studies*, **21**(4): 1487–1508.
Fabian, M. (2022), 'A psychologically-enriched version of tiberius' value-fulfilment theory of wellbeing', *Philosophical Psychology*, 1–25. https://doi.org/10.1080/09515089.2021.2016678.
Feldman, F. (2002), 'The good life: a defence of attitudinal hedonism', *Philosophy and Phenomenological Research*, **65**(3): 604–628.
Guala, F. and L. Mittone (2015), 'A political justification of nudging', *Review of Philosophy and Psychology*, **6**(3): 385–395.
Haidt, J. (2012), *The Righteous Mind: Why Good People are divided by Politics and Religion*. New York City: Penguin.
Harter, S. (2012), 'Emerging Self-Processes during Childhood and Adolescence', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York City: Guildford, 680–715.
Hausman, D. M. (2012), *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.
Hertwig, R. and T. Grüne-Yanoff (2017), 'Nudging and boosting: steering or empowering good decisions', *Perspectives on Psychological Science*, **12**(6): 973–986.
Higgins, T. (1987), 'Self-discrepancy theory: a theory relating self and affect', *Psychological Review*, **94**(3): 319–340.
Kahneman, D. (1999), 'Objective Happiness', in D. Kahneman, E. Diener, and N. Schwarz (eds), *Wellbeing: The Foundations of Hedonic Psychology*, New York City: Russell Sage Foundation, 3–25.
Kahneman, D. (2011), *Thinking: Fast and Slow*. New York City: Penguin.
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.
Kallbekken, S. and H. Sælen (2013), "Nudging' hotel guests to reduce food waste as a win–win environmental measure', *Economics Letters*, **119**(3): 325–327.
Kasser, T. and A. Ahuvia (2002), 'Materialistic values and well-being in business students', *European Journal of Social Psychology*, **32**(1): 137–146.
Kasser, T. and R. Ryan (1993), 'A dark side of the American dream: correlates of financial success as a central life aspiration', *Journal of Personality and Social Psychology*, **65**(2): 410–422.
Kasser, T. and R. Ryan (1996), 'Further examining the American dream: differential correlates of intrinsic and extrinsic goals', *Personality and Social Psychology Bulletin*, **22**(3): 280–287.
Kruglanski, A. and E. Szumowska (2020), 'Habitual behaviour is goal-driven', *Perspectives on Psychological Science*, **15**(5): 1256–1271.
Leary, M. (2012), 'Sociometer Theory', in P. Van Lange, A. Kruglanski and E. Higgins (eds), *Handbook of Theories of Social Psychology*, New York City: Sage, 151–159.
Leary, M. and R. Baumeister (2000), 'The nature and function of self-esteem: sociometer theory', *Advances in Experimental Social Psychology*, **32**(1): 1–62.
Lichtenstein, S. and P. Slovic (2006), *The Construction of Preference*. Cambridge: Cambridge University Press.
Martela, F. and K. Sheldon (2019), 'Clarifying the concept of well-being: psychological need satisfaction as the common core connecting eudaimonic and subjective well-being', *Review of General Psychology*, **23** (4): 458–474.
Nickerson, R. S. (1998), 'Confirmation bias: a ubiquitous phenomenon in many guises', *Review of General Psychology*, **2**(2): 175–220.
Novak, M. and R. Highfield (2011), *Supercooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. New York City: Free Press.
Paul, L. A. and C. R. Sunstein (2019), '"As judged by themselves": transformative experiences and endogenous preferences', *SSRN Electronic Journal*, https://dx.doi.org/10.2139/ssrn.3455421.
Paunov, Y., M. Wänke and T. Vogel (2019), 'Ethical defaults: which transparency components can increase the effectiveness of default nudges', *Social Influence*, **14**(3–4): 104–116.
Read, D. (2006), 'Which side are you on? The ethics of self-command', *Journal of Economic Psychology*, **27** (5): 681–693.
Rizzo, M. J. and G. Whitman (2019), *Escaping Paternalism: Rationality, Behavioral Economics, and Public Policy*. Cambridge: Cambridge University Press.

Ryan, R. and E. Deci (2004), 'Autonomy Is No Illusion: Self-Determination Theory and the Empirical Study of Authenticity, Awareness and Will', in J. Greenberg, S. Koole and T. Pyszczynski (eds), *Handbook of Experimental Existential Psychology*, New York City: Guildford, 431–448.

Ryan, R. and E. Deci (2017), *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York City: Guildford.

Schmidt, A. T. (2017), 'The power to nudge', *American Political Science Review*, **111**(2): 404–417.

Sen, A. (1985), 'Well-being, agency and freedom: the Dewey lectures', *The Journal of Philosophy*, **82**(4): 169–221.

Sheldon, K. M. and A. J. Elliot (1999), 'Goal striving, need satisfaction, and longitudinal well-being: the self-concordance model', *Journal of Personality and Social Psychology*, **76**(3): 482–497.

Sheldon, K. and M. Vansteenkiste (2005), 'Personal Goals and Time Travel: How Are Future Places Visited, and Is It Worth It?' in A. Strathman and J. Joireman (eds), *Understanding Behaviour in the Context of Time: Theory, Research and Application*, Mahwah: Erlbaum, 143–163.

Sheldon, K., N. Abad, Y. Ferguson, A. Gunz, L. Houser-Marko, C. Nichols and S. Lyubomirsky (2009), 'Persistent pursuit of need-satisfying goals leads to increased happiness: a 6-month experimental longitudinal study', *Motivation and Emotions*, **34**(1): 39–48.

Sheldon, K., A. Elliot, R. Ryan, V. Chirkov, Y. Kim, C. Wu, *et al*. (2004), 'Self-concordance and subjective well-being in four cultures', *Journal of Cross Cultural Psychology*, **35**(2): 209–223.

Showers, C. and V. Zeigler-Hill (2012), 'Organisation of Self-Knowledge: Features, Functions and Flexibility', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York City: Guildford, 105–123.

Silvia, P. and K. Eddington (2012), 'Self and Emotion', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York City: Guildford, 425–445.

Sobal, J. and B. Wansink (2007), 'Kitchenscapes, tablescapes, platescapes, and foodscapes: influences of microscale built environments on food intake', *Environment and Behaviour*, **39**(1): 124–142.

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

Sumner, L. (1996), *Welfare, Happiness and Ethics*. New York City: Oxford University Press.

Swann, W. (2011), 'Self-Verification Theory', in P. Van Lange, A. Kruglanski, and E. Tory-Higgins (eds), *Handbook of Theories of Social Psychology*, Volume **2**, New York City: Sage, 23–42.

Swann, W. and M. Buhrmester (2012), 'Self-Verification: The Search for Coherence', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York City: Guildford, 405–424.

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

Tiberius, V. (2018), *Well-Being as Value-Fulfilment: How We Can Help Each Other to Live Well*. New York City: Oxford University Press.

Vansteenkiste, M., B. Duriez, J. Simons and B. Soenens (2006), 'Materialistic values and well-being among business students: further evidence of their detrimental effects', *Journal of Applied Social Psychology*, **36**(12): 2892–2908.

Wallace, H. and D. Tice (2012), 'Reflect Appraisal through a 21st-Century Looking Glass', in M. Leary and J. Tangney (eds), *Handbook of Self and Identity: Second Edition*, New York City: Guildford, 124–140.

Woodman, T. and S. Hemmings (2008), 'Body image self-discrepancies and affect: exploring the feared body self', *Self and Identity*, **7**(4): 413–429.

Cite this article: Fabian M, Dold M (2025). Agentic preferences: a foundation for nudging when preferences are endogenous. *Behavioural Public Policy* **9**, 249–269. https://doi.org/10.1017/bpp.2022.17