# Individual Agency in Behavioral Public Policy: New Knowledge Problems

Malte Dold
*Pomona College*

**Abstract:** Taking psychology seriously in behavioral public policy (BPP) means to acknowledge the context-dependent and dynamic nature of people's preferences. Preferences adapt to situational context factors and sociocultural circumstances. Moreover, people often do not have given preferences but form them in the act of choosing. Insights on passive and active preference change pose a serious problem to the idea of nudging in BPP since it questions the policy analysts' ability to detect people's well-integrated 'true' preferences. Given this epistemic difficulty, some authors have recently begun to argue for agency-enhancing interventions in BPP, such as assistive cuing and boosts. Unlike nudges that re-bias people, these psychologically informed policy interventions aim at de-biasing people by enhancing their reasoning capacities. The agency-centric approach to BPP is laudable from a methodological perspective since it addresses intricacies of context-dependent preferences. It is also laudable from a normative perspective since it takes the liberal idea of individual self-determination seriously. Yet, it creates 'new' knowledge problems that have not yet been sufficiently addressed in the literature. This article argues that the epistemic challenge of an agency-centric BPP stems from, among other things, (a) its commitment to algorithmic analysis that models actual decision-making processes (in contrast to a reliance on algebraic analysis and as-if models) and (b) the difficulty to differentiate motivational from epistemic concerns that would allow the analyst to identify reasoning failures. This article discusses these challenges and ultimately defends a psychologically grounded agency-oriented approach to BPP that highlights the value of autonomy, i.e., people's capacity to scrutinize, act on, and identify with their evolving preferences. Such an approach attenuates some of the epistemic problems by shifting the policy focus from process facilitation and situational choice architecture toward institutional analysis.

**Keywords:** agency, behavioral public policy, boosting, context-dependence, knowledge problem

> "There is perhaps no 'correct' preference to have that policy should support. Nevertheless, it still matters (or ought to) for those with liberal instincts that whatever action people take, they should feel they own it in the sense that they have had the resources to reflect on what preferences to hold and how to act on them. That is, they should feel autonomous."
>
> Hargreaves Heap (2013, 995)

## 1. Introduction

Behavioral Public Policy (BPP) is a young and developing field. It is generally understood as the application of insights from behavioral economics and psychology to public policy analysis and design. A core insight from behavioral economics and psychology is that people's preferences are *context-dependent* (Sugden 2018). Context-dependence means that preferences are unstable and contextual features – such as informational framing, the arrangement of options, or defaults – influence people's choices. In many cases, people do not come to decision problems with a well-integrated preference ranking in their mind. Instead, they construct them 'on the go' in the process of choosing based on salient cues in the decision environment (Lichtenstein and Slovic 2006). Those cues become particularly relevant in cases of unfamiliarity, uncertainty, and complexity (KcKenzie et al. 2018). Context-dependence implies that people's preferences are often *time inconsistent*. Dynamic inconsistencies are not mere noise; they are the norm rather than the exception and emerge because of the inherently ad hoc, improvised nature of human thought (Chater 2022).

Dynamic inconsistency poses as serious problem for standard approaches in policy evaluation, such as Pareto or Kaldor-Hicks efficiency analyses (Cowen 1993). It also challenges the normative foundation of behavioral paternalism in BPP, in particular the idea of nudging. Nudging assumes that people are biased and impaired by a psychological 'shell' when making choices, e.g., this shell makes them pay too much attention to supposedly irrelevant features of the decision environment. Yet, individuals are supposed to be constituted by a core of well-integrated 'true' preferences that hold across different decision environments (Sugden 2018). In nudging policies, choice architects aim to redesign the decision environment to steer people toward their 'true' preferences without forbidding any options or significantly changing their economic incentives. However, if preferences are dynamically inconsistent, it is not clear at what point in time people's preferences should be a proxy for welfare (Read 2006). It is difficult to identify a well-integrated core of 'true' preferences that can be taken as a reliable benchmark for welfare-enhancing nudge interventions (Infante et al. 2016; Dold 2018; Sugden 2018; Rizzo and Whitman 2020).

The observation that preferences are context-dependent and often time inconsistent has recently motivated a number of scholars to move away from preferences as the yardstick for BPP to alternative normative criteria. For instance, Robert Sugden has proposed the *opportunity criterion* that suggests that "any expansion of a person's opportunity set promotes her interests, irrespective of her actual preferences" (Sugden 2018: 99). The idea is that people wish to obtain larger opportunity sets since they allow them to satisfy whatever preference they might hold in the future. This article will not discuss Sugden's insightful approach since it has been covered extensively in the literature (e.g., Vromen and Aydinonat 2021; De Rosa et al. 2021). Instead, it will discuss the proposal to focus on *individual agency* as an alternative benchmark in BPP.

Agency-centric approaches have recently been proposed by several scholars (Banerjee et al. 2023; Dold and Lewis 2023; Dold and Stanton 2021; Hargreaves Heap 2017, 2020; Hertwig and Grüne-Yanoff 2016, 2017; McKenzie et al. 2018; Schubert, 2015, 2021). While these approaches differ in how they conceptualize agency, they are united by their critique of BPP approaches that (a) assume behavioral outcomes as target variables (e.g., save more, eat less sugar, work out more, etc.) and (b) exploit citizens' cognitive biases as means to realize those behavioral outcomes. In contrast to nudges that typically exploit people's biases, agency-centric approaches aim to enhance *the quality of the cognitive process* that precedes choice by tools such as *assistive cues* or *boosts* that target people's competences (Hertwig and Grüne-Yanoff 2017). Unlike nudges that aim at a distinct behavioral outcome by re-biasing people, these psychologically informed policy interventions are process-oriented and aim at de-biasing people by enhancing their reasoning capacities.

The agency-centric approach to BPP is commendable from a methodological perspective since it explicitly acknowledges the context-dependent nature of preferences. It also circumvents the need to search for 'true' preferences by focusing on the decision-making process instead of behavioral outcomes. Moreover, it takes psychology seriously in that it acknowledges people's sense of agency as a key source of self-understanding and individual well-being (Dold et al. 2022). It is also laudable from a normative perspective since it takes the liberal idea of individual self-determination seriously. Yet, this article argues that agency-centric approaches create a set of new knowledge problems for policy-making that have attracted little attention so far. This article will argue that the epistemic challenges in agency-centric approaches originate from, among other things, (a) their commitment to algorithmic analysis that models actual decision-making processes (in contrast to behavioral economics' reliance on as-if models and algebraic analysis) and (b) the difficulty to differentiate motivational from epistemic concerns that would allow the analyst to reliably identify reasoning failures. This article will discuss these challenges and ultimately defend a psychologically grounded agency-oriented approach to BPP that highlights the value of autonomy, i.e., people's capacity to scrutinize, act on, and fully identify with their evolving preferences. Such an approach attenuates some of the aforementioned epistemic problems by shifting the policy focus from process facilitation and situational choice architecture toward institutional analysis. The wider agency perspective defended in this article provides a broad evaluative framework for BPP – a theoretical lens that helps analysts investigate the preference shaping power of norms, rules, and institutions and the crucial role of resources for individual processes of self-determination.

## 2. Agency in BPP

Agency-centric approaches acknowledge behavioral insights on the preference shaping power of the context but question the psychological realism of an 'inner rational agent' with stable preference that interacts with the world through an error-inducing 'shell.' Instead, they accept preference inconsistencies as part of the dynamic and becoming nature of humans. Consistency violations do neither necessarily lead to negative welfare effects (Arkes et al. 2018) nor do they reveal shortcomings in decision making. What appears as inconsistency is often part of an individual process of belief updating and preference learning (McKenzie et al. 2018). Getting rid of them can have negative effects on people's self-regulatory capacities (Rizzo and Whitman 2020).

### *Agency-enhancing interventions as process facilitation*

To differentiate between nudging and agency-centric interventions in BPP, it is helpful to divide the choice process into the choice *outcome* and the decision *process* leading up to it (see Figure 1). Nudging aims to facilitate certain behavioral outcomes (e.g., save more, eat less sugar, work out more, etc.). Typically, nudging exploits citizens' cognitive biases as means to effectuate behavioral change. They typically do not provide people with reasons to choose the nudged option but exploit the same flaws in deliberation that led to the 'biased' choice. A paradigmatic example would be a nudge that changes the default in retirement savings plans. Such a nudge does not enhance people's agency by helping them to reflect on their preferences. Rather, it takes advantage of people's tendency to stick to the status quo which directs them toward higher savings rates. The individual being nudged may not be conscious of the intervention or understand its intention. The choice architects restructure the decision-making environment 'top-down,' with the aim of guiding individuals towards the option that the architect deems to be in the best interest of the nudgee. In doing so, nudges "leave citizens 'in the dark', incapable of internalizing and owning the process of behavior change" (Banerjee et al. 2023, 1).

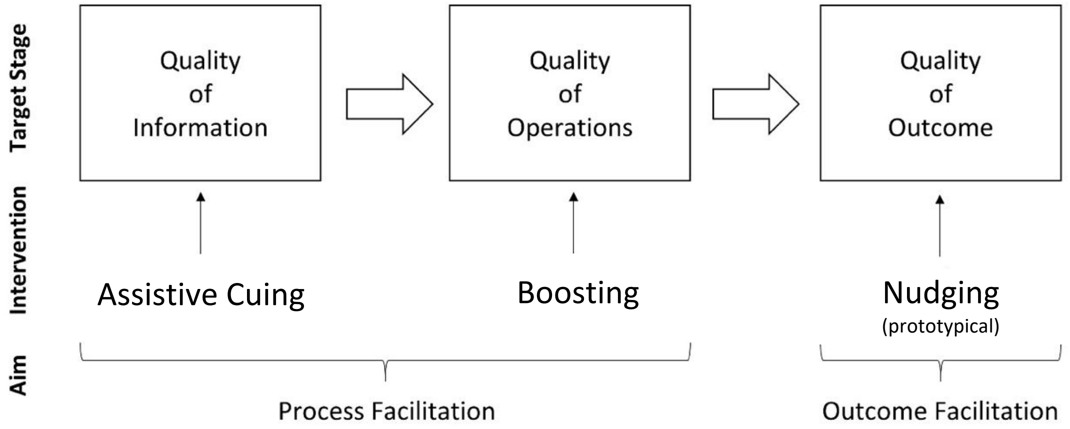

**Figure 1:** Process Facilitation (adjusted from McKenzie et al. 2018).

In contrast to nudging's focus on behavior change, agency-centric interventions aim at *process facilitation*. The goal is to enhance individuals' control of and identification with their choices and behavior change (if desired) by enhancing the quality of the decision process that precedes choice. By focusing on the process, agency-centric policy aims to "avoid taking a view on people's true interests while still being active if it was directed at the conditions under which people acquire the sense of interest on which they act." (Hargreaves Heap 2013, 995). Agency-centric policies are built on the insights that the quality of a decision process is jointly determined by (1) the quality of the information fed into the process and (2) the appropriateness of the operations that transform information into option-selections (McKenzie et al. 2018). In the following, we will discuss *assistive cuing* as an example for (1) and *boosting* as a prominent example for (2).

*Assistive cuing*

In many situations, people look to the context for relevant cues since they face a choice for which they lack clear antecedent preferences or sufficient knowledge about the various attributes of options. Proponents of assistive cuing argue that "[this] view opens up an alternative role for the choice architect – not as nudging parent, but as cooperative communicator, crafting contexts that effectively convey valid and useful information to decisionmakers." (Sher et al. 2022, 522). *Assistive cues* are interventions that alter the choice environment by providing 'high quality' information to facilitates accurate choice-relevant inferences. The goal of assistive cueing is to make reasoning processes easier by constructing representative choice environments that preserve the statistical structure of the relevant natural environment (McKenzie et al. 2018).

In difficult choice environments for which people have limited prior knowledge of the distribution of product attributes, the options sampled in the choice menu may lead them to update their beliefs about the market.  For example, if first term house buyers are unfamiliar with the interest rates for 30-year loans, they may use the mean value M of their sampled menu (let's say they received quotes on interest rates from five banks) to estimate the likely mean value for interest rates in the market. These sample-based inferences may in turn affect people's preferences: if M is relatively high, people are willing to pay more for the loan and if M is low, their willingness-to-pay is lower; of course, this means that their WTP depends on the sample which might not accurately reflect the mean value of interest rates in the market. To help people make correct inferences about the mean value in the market, assistive cuing would design a sample whose mean value accurately reflects the mean value of the market.

Another example are judgments made in joint vs. separate evaluation scenarios that trigger different evaluation processes. For instance, in a screening situation of candidates whose respective qualities differ along multiple attributes, the evaluation of one candidate in isolation ({A} or {B}) or both candidates jointly ({A, B}) leads to different estimates of their qualities. In joint evaluations, attributes that are unfamiliar typically receive greater weight since the joint view allow for a better assessment of the meaning of the unfamiliar attribute. As an illustration, in a study conducted by Hsee (1996), participants assess either one or two candidates for a programming job that involves a made-up computer programming language. Candidate A with a higher GPA, a widely known attribute, received a higher average salary from subjects when evaluated alone. In contrast, candidate B has a lower GPA but more experience (she produced a larger number of programs in the past, a highly relevant but less familiar attribute). Candidate B receives a higher salary when evaluated alongside candidate A. Apparently, the joint evaluation menu provides more information and likely leads to a more accurate assessment since it allows agents to compare the relative sizes of various attributes of the job candidates. Assistive cuing would therefore try to provide agents with the joint evaluation set when evaluation sets are representative (McKenzie et al. 2018).

### *Boosting*

Boosts are another type of agency-enhancing interventions that aim at process facilitation. Boosts are explicitly intended to "preserve and enable individuals' personal agency." (Hertwig and Grüne-Yanoff 2017, 982). Boosts aim at expanding the decision-makers' competence to reach their own objectives, “without making undue assumptions about what those objectives are” (Grüne-Yanoff & Hertwig 2016, 156). Unlike assistive cues that aim to enhance agency by improving the quality of the informational input fed into the existing decision process, boosts target agents’ repertoire of cognitive strategies that transform information into choice (McKenzie 2018 et al., 356). They do so by training them in the use of helpful decision heuristics. These heuristics are simplifying rules of thumb that help people achieve their goals in certain types of environments. Ideally, employing these decision heuristics is better for the achievement of a given purpose than the cognitive strategies people currently rely on. And unlike nudges or assistive cuing, boosts “require the individual’s active cooperation” and that “[i]ndividuals choose to engage or not to engage with a boost” (Hertwig and Grüne-Yanoff 2017, 982). Boost interventions can only be effective if the person being boosted accepts the training, internalizes the necessary competence, and utilizes it in the right moment when faced with a problem. According to Grüne-Yanoff (2018) these three factors make sure that any change in behavior resulting from a boost is grounded in reason.

The boost literature has produced an impressive list of policies (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017). At least three types of policies can be distinguished within the boost approach: (1) boosts that foster *risk competence* when risks are known and measurable; (2) boosts that foster people’s *domain-specific competence* by teaching smart behavioral rules of thumb; and (3) boosts that teach *fast-and-frugal decision trees* in situations of uncertainty.

An example for (1) is teaching people how to translate one risk format into another, viz., relative probabilities into natural frequencies (Hertwig and Grüne-Yanoff 2017, 977). This boost is motivated by findings that people systematically neglect base rates in Bayesian inferences. People tend to ignore base rates in favor of salient individuating information, rather than integrate the two. This important implications for judgments in many clinical, legal, and social-psychological settings. The base rate neglect largely disappears when subjects are taught with a computerized tutorial program to use a simple heuristic that helps them construct frequency from probability representations by means of frequency grids or frequency trees (Sedlmeier and Gigerenzer 2001). This representation training has a higher immediate learning effect as well as greater temporal stability than explicit trainings of how to insert probabilities into Bayes’ theorem.

An example for (2) is training people in temptation bundling (Milkman et al. 2013). A simple behavioral rule of thumb can help people overcome self-control problems by bundling ‘want’ experiences that are instantly gratifying (e.g., listening to desirable audiobooks or podcasts) with valuable ‘should’ behaviors that provide delayed rewards (e.g., exercising or doing the dishes). For this temptation bundling to work, people need to allow themselves to enjoy the ‘want’ activities only while doing the ‘should’ activities, thus making the latter more tempting. Evidence in the context of gym attendance suggests that such activity bundling leads to higher uptakes of ‘should’ behavior: In comparison to a control group, the subjects using temptation bundling showed a higher weekly gym attendance rate (0.94 vs. 1.42 visits/week). Temptation bundling is a boost since it enriches people’s cognitive repertoire by training them to use a simple decision rule (‘Whenever I *should* do something, combine it with something I *want* to do.’). This heuristic helps people interpret and navigate their environment in novel ways (‘What are ‘should’ and what are ‘want’ activities and how can I combine them fruitfully?’) (Grüne-Yanoff et al. 2018, 253).

An example for (3) are fast-and-frugal decision trees (FFTs). FFTs are a special kind of decision tree with at least one end node after each decision node. They offer several advantages over full trees: they require less information, have a simple decision rule, and offer robust predictions in situations of low-quality data. An example of a FFT is a diagnostic technique that enables quick medical diagnoses based on a few physical cues. For example, Marewski and Gigerenzer (2022) argue that a tool to predict and treat pneumonia in young children can be designed as a simple decision tree based on two cues: duration of fever and a child’s age. Similar techniques can help laypeople to identify signs of a heart attack or stroke and prompt them to seek timely assistance (Grüne-Yanoff and Hertwig, 2016).

## 3. New Knowledge Problems

There are several normative and practical advantages of agency enhancing interventions compared to more ‘traditional’ BPP interventions such as nudging (Banerjee et al. 2023). Agency-enhancing interventions are less manipulative since they address people’s reasoning capacities and do not exploit their systematic biases. Particularly in the case of boosts, they are more transparent since they require the buy-in of the affected citizens. Since the interventions are focused on information giving (in the case of assistive cuing) and on competence building (in the case of boosts), psychological reactance (‘backfiring effect’) against them will likely be lower. Moreover, building up an informational stock and decision competences may contribute to non-persistent treatment effects since people internalize the process of behavior change. And while both policies reduce the epistemic burden in BPP relative to nudges as they do not demand knowledge about people’s true preferences, they pose several ‘new’ knowledge problems. This section will address them and ask what type of knowledge policy analysts would need to acquire to implement assistive cues and boosts effectively and legitimately. The discussion in this section does not attempt to be exhaustive but is instead meant as a conversation starter. While some of the knowledge ‘hurdles’ mentioned here have been acknowledged by proponents of agency-enhancing interventions, others have not yet been sufficiently addressed in the BPP literature.

*Challenge I: Understanding actual reasoning processes*

The nudging approach is built on the heuristics-and-biases approach (HB) which relies on parametric extensions of MaxU models in a world of risk (Gigerenzer, forthcoming). A paradigmatic example is the prominent beta-delta model of quasi-hyperbolic discounting that models intertemporal choice: $U^0(u) = u^0 + \sum_{i=1}^{\infty} \beta\delta^i u^i$ where $\delta$ captures the individual's 'standard' time preference of exponential discounting and $\beta$ captures the parametric extension that captures individual's 'behavioral' impulsivity. This approach follows the as-if logic of standard economics. Most behavioral economists agree that the beta-delta model doesn't capture actual decision-making processes, i.e., it doesn't entail the claim that people reach their decisions by maximizing their utility by discounting the future exponentially ($\delta^i$) but then get sidetracked by an additional factor ($\beta$) that discounts the next moment in time more heavily. Instead, as in standard economic methodology, they believe that the model simply summarizes observed behavior following the idea that people behave *as if* they calculated their utilities by means of the quasi-hyperbolic function. This is a functionalist-dispositional account (Grüne-Yanoff 2022): the functional mapping of the inputs (the features of the choice task, including the discount factors, the alternatives, and the respective outcomes) to outputs (the chosen alternative) matter. If a given function cannot perform a mapping that fits the observed data, then the functional form needs to be revised, e.g., by including more or different parameters.

In contrast, agency-enhancing tools such as assistive cues and boosts are built on the simple heuristics framework (SH) that aims to understand the process of *actual* decision-making in situations of uncertainty and study how minds interact with their environment. As McKenzie et al. (2018, 350) state: "Before we can improve people's decision-making behavior, we must first understand it." While this more ambitious approach is commendable on several grounds, e.g., for explanatory and predictive purposes, it is also a taller epistemic order and requires a more nuanced and fine-grained analysis of decision-making in different contexts. The SH approach aims to investigates the repertoire of heuristics people actually use to solve problems *and* the environment under which each heuristic is successful. In doing so, SH is not a deductive-axiomatic approach but an empirical investigation of the cognitive repertoire of heuristics and their respective ecological fitness. The crucial question is how to identify "the right rule of thumb for the right situation" (Gigerenzer 2007, 49). Identifying the 'right' heuristics and studying its applicability or selection criteria can be a lengthy research process. For instance, take the aforementioned and now established finding of the effectiveness of relative frequency representations. Decades of research went into first identifying base rate neglects in the late 1970s, then realizing in the 1990s that the mind's statistical reasoning processes evolved to operate on natural frequencies and that people can be effectively taught how to translate probabilities into frequencies to simplify the Bayesian computations. And finally, in the 2000s, researcher studied how boosts can be effectively taught to help people in practical diagnostic situations (e.g., by teaching medical and legal professionals about the effectiveness of frequency trees).

Conceptually, the higher relative epistemic burden of the SH framework stems from the need to acquire richer mechanistic evidence.[1] This can be captured by contrasting stylized mechanistic models for the two main psychological approaches in behavioral public policy, the HB and the SH approach (Grüne-Yanoff et al. 2018). While the following is a simplified representation that omits certain features of a realistic description, the mechanisms of heuristic decision-making involve the following components: (1) the *heuristic repertoire R*, which refers to the set of heuristics individuals has access to, (2) the *environment E,* which is typically referred to as the choice context, and (3) the resulting *behavior B.* The HB approach assumes a stable relationship between heuristics and environment: the type of environment $E$ always triggers the same heuristics and hence leads to predictable biases; changes in environment $E$ – for instance in the form of nudge*s* – lead to predictable changes in behavior $B$*:* $B = h(E)$ with $h$ being the heuristic function that represents the causal influence of $E$ on $B$*.* In contrast, in the SH framework the relationship between environment and heuristics is *dynamic*. The chooser selects heuristics adaptively depending on their environmental fit. Behavior cannot be predicted by studying heuristics in isolation but must be seen as the outcome of an interplay between the person's cognitive repertoire and the environment: $B = h(R, E)$. In the simple heuristics framework, policy designers cannot assume the stability of a heuristic in a given environment but need to collect data about the complex interaction between environment and cognitive repertoire of heuristics ($E$, $R$) that together result in behavioral outcomes. This is further complicated by the fact that the behavioral outcome $B$ is mediated by the agent's – deliberate or unconscious – selection of heuristics from $R$*.* Accordingly, behavior can change even when the environment $E$ does not change. This is unusual for economists because it means analysts cannot just study changes in the incentive structure of environment to explain behavior change but need to study 'what is going on in people's minds' to explain and predict their behavior. It is also unusual for behavioral economists, since they have traditional focused on the cognitive side while neglecting the interaction of cognitive processes with the environment (Gigerenzer, forthcoming). Admittedly, in many ways this higher relative epistemic burden of the

[1] In the context of BPP, the term mechanistic evidence has been coined by Grüne-Yanoff (2016). Mechanisms describes the causal process which, given certain circumstances, predictably produce one or more effects. Accordingly, mechanistic evidence in BPP is information concerning the causal pathway connecting a behavioral intervention to its outcome.

SH framework is not a bug, but a feature that follows from its commitment to a more detailed investigation of decision-making processes and its goal to match those cognitive processes with environmental structures.

Applied to the two agency-enhancing polices discussed above, in the case of boosting, policy analysts need to acquire knowledge of the *actual* cognitive heuristics repertoire of the human mind, identify under which environmental conditions these heuristics work and under which ones they are misleading, and then convey this information in the form of teachable lessons to people to enhance their decision-making competence (Grüne-Yanoff and Hertwig 2016). Simple heuristics may work well in certain environments. However, not all environments have a structure that can be successfully matched by simple decision rules. Some environments may require more computationally intensive strategies for satisfactory decision-making. In the case of assistive cueing, policy designers need to acquire an understanding of the signals that environmental cues convey (information leakage) and the specific cognitive heuristic they trigger (information absorption).[2] The critical question is how informative or misleading the information in the sampled choice set is, i.e., how closely does the attribute distribution in the sample resemble the true distribution in the natural environment (McKenzie et al. 2018). Moreover, both boosts and assistive cues require an understanding of the dynamics of attention. Belief-updating and competence building must conform to people's cognitive capacity limits of attention and working memory (Marois and Ivanoff 2005). The provision of explicit or implicit information as well as the teaching of competences will only be effective if analysts possess knowledge of the structural boundaries of people's attention, motivation, and time. To effectively communicate with decision-makers, analysts must follow two principles: Calibrate the amount of information presented to decision-makers' attentional capacity; and, when selecting and organizing information, ensure that salient information is relevant information (McKenzie 2018, 354; Sher et al. 2022, 522).

These are all technical questions that require the integrated knowledge of psychologists, cognitive scientists, and 'field specialists,' such as economists who have knowledge of the distribution of attributes in the natural environment. The interplay between the mind and the decision context is complex. Due to the multi-faceted nature of human cognition context effects can arise in myriad ways, and there is a latent danger in both policies that analysts might miscalibrate some of their policies, particularly in a dynamically changing environment. In the case of assistive cues this means that policy analysts need to check whether the sampled evaluative choice set still reflects

[2] For a discussion of information leakage and information absorption in public policy contexts, see Krijnen et al. (2017).

the statistical structure of the environment. And in the case of boosts analysts need to check whether the new environments still have a cue structure that can be matched successfully by some simple rule.[3]

*Challenge II: Identification of process breakdowns*

To devise and implement agency-enhancing policies, policy designers need to identify domains in which people experience difficulties in reasoning, e.g., due to knowledge gaps, poor quality of choice set, problems in statistical or risk assessments (Grüne-Yanoff and Hertwig 2016, 167). This means that they need to find out "what aspects of decision making are in need of improvement" (McKenzie et al. 2018, 351). To identify those aspects, policy analysts need to specify a notion of agency-reducing epistemic errors. To illustrate this point, Grüne-Yanoff (2022) differentiates between the *algebraic* and the *algorithmic* level of analysis. The functionalist-dispositional account is an algebraic level analysis: for instance, intertemporal decision-making is understood as a computational 'problem' and observed behavior is summarized by the quasi-hyperbolic functional form. In contrast, the process analysis of the SH framework proceeds on the algorithmic level in that it tries to capture the mental steps (including memory retrieval, imagination, mental representation, etc.) agents actually employ to arrive at a certain decision.

Grüne-Yanoff (2022) argues that a shift to the algorithmic level can help the analyst separate motivating from epistemic factors and this can in turn help detect epistemic errors in decision-making processes. For instance, people's tendency to discount the future – as captured by the beta and delta factors in the context of quasi-hyperbolic discounting, have been interpreted as entailing an epistemic and a motivational component. On the one hand, time discounting is commonly perceived to be *motivated* by the uncertainty of human life and a psychological discomfort of deferring available gratifications (Frederick et al. 2002). Yet, discounting can also entail an *epistemic* component when it is driven by people's lack of ability to imagine their older selves. In a famous passage from *The Economics of Welfare*, Pigou (1920) suggested that discounting is a type of cognitive illusion: "our telescopic faculty is defective, and we, therefore, see future

[3] This point can be illustrated by the recognition heuristic which helps people estimate an unknown criterion (e.g., the population of a foreign city) by relying on a simple rule: If one object is recognized and the other is not, it can be inferred that the recognized object has the higher value with respect to the criterion. For instance, if people recognize the name of a foreign city on a list while not recognizing others, then this option is likely to rank highest in terms of population and choosing it will likely be successful. However, the recognition heuristic breaks down if the reason why people recognize a particular option is no longer valid (e.g., the population of the respective city decreased due to some exogenous shock people are not aware of).

pleasures, as it were, on a diminished scale."[4] Modern psychology and behavioral economics tends to support Pigou's epistemic interpretation of time preferences. For instance, Hershfield et al. (2011) presents subjects with AI renderings of their future selves. Subjects who are exposed to their virtual future selves presumably overcome their limited imagination and discount the future less severely. If the study found that that subjects still discounted the future heavily even when they are confronted with their older selves, it would be less likely that their discounting is based on an epistemic distortion but rather should be interpreted as a subjective evaluation of the uncertainty that necessarily involves future states. In another study, Loewenstein and Prelec (1993) demonstrates how assistive cuing in the form of choice bracketing can influence people's discount rates. The authors present one group of subjects with a smaller choice set ({French restaurant in 1 month; French restaurant in 2 months}); the other group is exposed to an informationally richer choice set ({French restaurant in 1 month *and* eating at home in 2 months; French restraint in 2 months *and* eating at home in 1 month}). In first case of the informationally poorer environment, most subjects choose the French restaurant in one month, presumably reflecting their epistemic error to perceive all available options. In the second case, a majority opts for the French dinner in two months, presumably reflecting people's patience and motivation for improvement sequences (Frederick et al. 2002, 375).

If such experimental evidence reliably revealed that time discounting involves distorted cognitive or perceptual elements on behalf of decision-makers, algorithmic analysis would have helped separate motivating from epistemic factors and identified cognitive errors in time discounting. Yet, this approach poses a high epistemic burden for policy designers.[5] Policy analysts would need to be able to collect sufficient evidence of detailed cognitive processes to differentiate with high confidence whether discounting represents people's subjective evaluation of future uncertainty *or* arises from people's limited ability to imagine future states. This differentiation is notoriously difficult since both aspects (and other confounding factors) play likely a role in people's tendency to discount the future (Frederick et al. 2022, 380). Moreover, analysts typically do not have direct

[4] Sunstein (2020, 202) highlights the epistemic component in intertemporal choice: "Recall that choosers must solve a prediction problem; they must decide, at some point in advance of actual experience, about the effects of one or another option on experience. To solve that problem, knowing 'how they feel' is not enough. At a minimum, they must know 'how they will feel,' and they might not know enough to know that."

[5] A point that is readily admitted by Grüne-Yanoff (2022, 161-2): "Success of such investigation depends on a number of assumptions – (a) that there is a fact of the matter about the agent's representations and algorithmic steps; (b) that these facts are sufficiently stable that they are generalizable across contexts and agents; and (c) there are publicly observable factors on which an experimenter can intervene in order to infer posited representations and algorithmic steps. All of these assumptions can only be empirically validated."

access to cognitive processes but must rely on behavioral evidence elicited in choice situations under controlled conditions to 'build up' their algorithmic analysis. The epistemic demand of this approach is further increased by the fact that it is difficult to differentiate between intrinsic motivation and derived motivation (Grüne-Yanoff 2022, 162). Intrinsic motivation (e.g., to save for retirement) is motivation that cannot be judged to be erroneous without violating the principle of subjectivity. In contrast, derived motivation (e.g., to invest in stocks to save for retirement) is computed from intrinsic motivation and includes epistemic components (e.g., how volatile are stocks; what is their relative return of investment, etc.). While this search for intrinsic motivations can be grounded in psychological theory (and is thus not prone to the critique of psychological unrealism, see Infante et al. 2016), doubts are warranted whether policy designer can indeed gain access to temporally stable intrinsic motivations given the context dependence of cognitive processes.

Admittedly, in the case of boost interventions policy designers need to require only minimum knowledge of intrinsically motivated goals since the identification is typically "general and conjectural" (Grüne-Yanoff 2021, 299): it suffices to argue that some people in some situations hold such intrinsically motivated goals.[6] Unlike in the nudging case, policy designers do not need to know how intrinsically motivated goals and epistemic mistakes are distributed in the population since boost policies typically leave the choice of boosts to people themselves (e.g., it remains the decision of individuals whether they apply temptation bundling to overcome self-control or apply simple accounting rules). Still the argument of this section holds that they need to acquire *some* knowledge of intrinsically motivated goals to start their analysis and propose the 'right' type of boosts or assistive cues. Such knowledge is easier to come by the more general the assumed goal is (e.g., statistical literacy and accurate risk assessment); yet it might be trickier in more personal cases (e.g., effective self-control in dietary decisions) where intrinsic goals are harder to identify due to the complex nature of intertemporal choice and the cognitive processes involved.

This latter point has not yet been sufficiently addressed in the literature on *motivational boosts* whose goal it is to foster the competence to adjust one's self-control through interventions such as attention and psychological connectedness training or reward-bundling exercises (Hertwig and Grüne-Yanoff 2017, 979). In these cases, the knowledge problem is particularly intricate since proposed motivational interventions are transformative by nature and will likely affect people's

[6] Grüne-Yanoff (2021, 300): "The boosting policymaker instead only has to ensure that for the population on average, such a boost would be useful, desirable, and not undermining; and such general pattern knowledge can typically be acquired through standard social science research."

intrinsic motivations. Interventions will often change people' habits and routines (e.g., healthier eating) and may hence lead to durable changes in intrinsic motivation that would not have happened without the intervention (Fabian and Dold 2022). Admittedly, many of these motivational changes might be embraced by people (after all they are the ones who actively applied the boosts). Yet, it is conceivable that in some cases these motivation changes happen in the form of subtle processes and as reactions to environmental stimuli people are neither fully aware of nor would embrace as their own if they consciously reflected on them. If boosts shape the process of constructing preferences in such subtle ways, they fail to fully endorse the idea of agent goal subjectivity.

## 4. Widening the view: Agency concerns beyond process facilitation

What are implications from the discussion of knowledge problems resulting from the complex interaction between people's minds and their changing environments, the difficult differentiation of motivational from epistemic concerns, and the intricate issue that preferences might be shaped by trying to improve the choice process? One possibility is to double down on the study of cognitive processes in the hope that analysts acquire more fine-grained knowledge that allows them to design and implement interventions that would facilitate individuals' reasoning processes while making sure that the interventions only transform derived but not intrinsically motivated goals. Yet, the latent tension in this approach is that it might fall short of its high epistemic aspirations and, more importantly, miss a crucial aspect of people's agency concerns. Approaches such as assistive cuing and boosting focus on process facilitation. Their main goal is to rectify epistemic problems in means-ends calculations by changing either the situational environment (in the case of assistive cues) or the mental operations that translates situational cues into actions (in the case of boosting). In doing so, they aim to enhance the epistemic competence of the decision-maker. Clearly, being able to make competent decisions is crucial for agency, i.e., for one's sense of being the author of one's own life. In fact, psychological studies highlight that competence contributes to people's *sense of agency* since it enhances the experience of self-efficacy (Bandura 2006) as well as effectiveness and mastery (Ryan and Deci 2020). However, those studies also highlight that besides competence, autonomy is crucial, i.e., "the experience of volition and willingness … when one's actions, thoughts, and feelings are self-endorsed and authentic." (Ryan and Deci 2020, 3). Self-endorsed and authentic actions are actions "really proceeding from its reputed author" (Ryan and Deci 2006, 1561). Autonomy does not need to entail an absence of external influences: if a person concurs with or endorses acting in accord with an external demand, a person can still act authentically (Ryan and Deci 2006, 1560). Generally speaking, autonomy "obtains when, were one to turn a reflective eye toward the motives, values, and concepts that structure one's judgments (and do so in a piecemeal manner), one would not feel deep self-alienation." (Christman 2005, 345).[7] Of course, competence contributes to a person's autonomy, e.g., when consumers learn how to interpret statistics surrounding health or food risks, they are more likely to be able to make choices they fully endorse as their own; statistical illiteracy, on the other hand, would leave citizens vulnerable to ' 'techniques that deliberately and insidiously exploit limited statistical literacy in order to convince the audience that they are at high risk of illness' ' (Gigerenzer et al. 2007, 71). Yet, autonomy exceeds the ability to make correct inferences in means-ends calculations: making autonomous choices also means that one is able to reflect upon, act on, and identify with one's *evolving ends*.

*The self as a process*

The wider view of agency takes psychology seriously, in particular the idea that "the self is a process" (Sheldon 2022, 110), i.e., a reflexive, dynamic project that the individual 'works on' in a discursive interchange with others. It embraces the idea that "modern adults think of themselves as growing, changing, moving through on-time passages and stages, as the self forms a trajectory of development from the remembered past to the anticipated future" (McAdams 1996). Such developmental thinking provides the empirical background for the conceptual widening of the notion of agency beyond concerns for decision competence.

There is a rich literature at the intersection of philosophy and economics highlighting that many choices are *transformative experiences* that people seek out knowing they will affect their preferences, such as travelling, having a child, entering a relationship, enrolling in university, or going to therapy (Callard 2018; Roberts 2022). The idea of active preference change has been advocated by liberal political economists from Adam Smith to Frank Knight to James Buchanan (Dold and Rizzo 2021, Matson 2022). According to these thinkers, agency involves a dynamic component: it is not simply choosing competently the right means to satisfy a static set of wants but transforming oneself to have different (and potentially better) wants. Of course, life choices do not have to be done with the intention to be transformative to influence our preferences in profound ways. Preference change is often simply the byproduct of experience, e.g., when consuming certain

[7] Christman (2005, 346) clarifies that the self-reflection involved in this test for autonomy "is purely subjective in that it takes as its perspectival orientation the agent's own point of view, independent of any external account of the motives, values, and beliefs to which she might turn her attention." Moreover, this sense of autonomy is not built on the idea of a 'true' self; rather it describes processes of self-reflective endorsement; what one endorses may well change over time (Christman 2009, 134).

food or music (Loewenstein and Angner 2003) or spending time at work or with friends where emulation and the exposure to social norms and rules shape our preferences (Frank 2021).[8] Having agency means to be able to become aware of and reflect upon those passive preference changes in order to choose those that one can fully endorse as one's own.

Problematic from an agency perspective are preferences changes that result from unconscious adaptive processes where people adjust their preferences to situations of economic destitution or social pressure (Elster [1985] 2016; Sen 1999). While one can justify such preference adaptations on incentive grounds in some cases (i.e., it might simply be too harmful for people to hold onto their 'old' preferences), adaptive preferences often persist even when incentive structure change since people have internalized the social norms into their belief system. In many cases, people still believe in the 'righteousness' of oppressive social norms, such as foot binding, female mutilation, or the caste system, even though there is no external punishment mechanism that forces them to do so (Hoff and Stiglitz 2016). In such cases, 'blind' rule-following behavior is problematic since the social norms influence the options perceived to be available by indicating what is 'permissible' (Sen 1997). Generally, much of people's rule-following behavior is habitual. In such cases, what people belief in and what they choose might fulfill the competence condition, yet it can in some cases violate the autonomy condition of agency.

### *An even taller order?*

Does this wider understanding of agency further increase the epistemic burden for policymaking? Not necessarily so. The acknowledgment of autonomy as a core ingredient for people's sense of agency might imply that policy analysts need to be less focused on concrete outcomes (as in the case of nudges) or the fine-grained analysis of *internal* cognitive processes (as in the case of assistive cuing and boost) and rather focus their analysis on the *external* social conditions (the norms, rules, institutions, and resources) that facilitate and constrain people's efforts to reflect, act upon, and identify with their evolving preferences. Given the ubiquity of social influences on people's preference and belief formation, people face the problem to discern among those influences what is meaningful to them and what withstands critical scrutiny. This is a process that policy can support through attention to the social conditions under which people make such reflections. A behavioural public policy that takes the wider view of agency seriously "would seem naturally to be concerned with the conditions (e.g., the educational system, the media, the family,

[8] The social life of work, family, and friends can be considered the main domains wherein selves are made and remade Taylor (1989, 209) calls this "the affirmation of the ordinary life."

vibrancy of the arts world) that support reflection on what preferences to hold" (Hargreaves Heap 2013, 995). The quality of those institutions could then be judged based on whether they allow people to choose life plans that they fully identify with "in the sense that they have had the resources to reflect on what preferences to hold and how to act on them" (ibid). Accordingly, an agency-enhancing BPP would study the institutional prerequisites for a dynamic society in which people have the chance to explore the impact of social influences and develop themselves through Millian 'experiments in living' (Delmotte and Dold 2021).

The focus on social conditions for individual agency is a point that Amartya Sen makes forcefully in his book *Development as Freedom*. Sen emphasizes that "[there] is a deep complementarity between individual agency and social arrangements. It is important to give simultaneous recognition to the centrality of individual freedom and to the force of social influences on the extent and reach of individual freedom." (Sen 1999, xii). Analysts committed to the study of 'the force of social influences' on individual agency would need to broaden the BPP discourse by studying social determinants of individual preference and belief formation processes. This would include the social environment to which people are exposed and have become accustomed to as well as the cultural mental models (including categories, identities, narratives, and worldviews) that they have internalized to process information (Hoff and Stiglitz 2016; Dold and Lewis 2022). Such a program would combine insights from psychology and sociology and likely pose new epistemic challenges for BPP whose discussion lies beyond the scope of this article.

## 5. Conclusion

Motivated by the challenge to identify people's context-independent 'true' preferences, recent approaches in BPP have argued for individual agency as a policy goal. Compared to nudging whose focus lies on behavioral *outcomes* and the exploitation of people's biases, agency-centric approaches appeal to people's reasoning capacities and focus on the quality of decision *processes*. This article discussed two prominent policy ideas of agency-centric BPP, assistive cuing and boosting. Assistive cues aim to facilitate accurate choice-relevant inferences. By providing inputs that are as clear, accurate, and relevant as possible, they strive to improve the quality of the informational input fed into existing decision processes. Boosts in contrast aim to help people develop decision-making competences for different domains of their lives by training them in the use of simple heuristics. They thus address the level of cognitive operations that transform information into option-selections. This article highlighted that both policies focus on epistemic factors in process facilitation and understand agency mainly as decision competence, i.e., correct means-end calculations. While both policies have advantages over the nudging approach, they still run into intricate knowledge problems that stem from their commitment to analyze actual decision-making processes (as compared to as-if models) and the need to differentiate between epistemic and motivational concerns when identifying decision mistakes. The latter point is particularly tricky in cases of transformative interventions, such as motivational boosts. Based on these epistemic concerns, the article discussed the potential of a wider notion of agency in BPP that stresses autonomy instead of competence, i.e., the idea that the central focus in agency discussions should be people's capacity to reflect on, act on, and identify with their evolving preferences. Competence still matters in this view, but mainly instrumentally for people's sense of self-authorship.

The epistemic advantage for applied BPP is that this wider approach focuses less on internal factors of agency (such as the quality of cognitive processes) and asks instead what are external social conditions (such as rules, norms, institutions, and resources) that are conducive to individual processes of self-constitution. The conceptual added value is that this wider approach highlights the pervasive role of social factors in individual processes of preference and belief formation. It places less emphasis on the question of whether the concrete context in which decisions are made is failure-reducing or not and rather asks whether the wider social context is autonomy-enhancing or not. It also widens the temporal dimension of agency analysis from understanding it as accurate instrumental reasoning at discrete moments in given environments to understanding it as self-endorsed processes of becoming whose quality depends on a series of choices people make in dynamic social environments. The guiding question changes from 'Do I feel competent in making choice X in environment Y?' to 'Do I feel that I am the author of my own life?' In this sense, agency is crucially a quality of people "living out a reflectively endorsed autobiographical narrative that reflects embedded values operative over time and across conditions" (Christman 2009, 160).

Any model of agency proposed in public policy is a thick concept: it describes and evaluates simultaneously (Alexandrova and Fabian 2022). The litmus test for such models should be whether individuals feel reflected in the proposed model. It is an open question whether this is the case for the 'narrower' model that emphasises competence, the 'wider' model that emphasizes autonomy, some combination of the two, or whether people do care at all about agency. By raising awareness of how social conditions shape people's beliefs and preferences, or by prompting individuals to reflect on their own decision-making errors, agency-centric policies may be perceived as intrusive and irritating. Ultimately, these are empirical matters and can only be clarified by including the affected citizens. In this regard, BPP can learn from recent discussions in well-being public policy (WPP) that emphasize that policy criteria should be co-produced in processes of public deliberation in mini-publics and other forms of organized public discourse.[9] Besides weighing in on which agency concept reflect citizens interest best, public deliberation can also help identify institutional preconditions and necessary resources for agency, particularly on the local level. In doing so, people are actively involved in shaping the social rules under which they want to live, and not relegated to passive recipients of the fruits of 'smart' behavioral interventions. This follows Chater (2022, 1) who has recently argued that behavioral insights "do not override, but can (among many other factors) inform, our collective decision-making process. The point of behavioral insights in public policy is primarily to inform and enrich public debate when deciding the rules by which we should like to live." Academic expertise can enrich public deliberation by helping citizen and politicians better understand the social conditions for and challenges of individual agency. Of course, public deliberation is not without problems. It can strengthen or create problematic features of decision-making (e.g., motivated reasoning, herd behavior, group think, etc.). Yet, if accompanied by the right, inclusive rules of public discourse such deliberative processes can help citizens form and share their beliefs about agency-centric BPP. The possibilities and limits of public reasoning in the context of agency as a policy goal should be addressed in future research.

## References

Alexandrova, A., & Fabian, M. (2022). Democratising measurement: Or why thick concepts call for coproduction. European Journal for Philosophy of Science, 12(1), 7.

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision, 3,* 20-39.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science, 1*(2), 164-180.

Banerjee, S., Grüne-Yanoff, T., John, P., & Moseley, A. (2023). It's Time We Put Agency into Behavioural Public Policy. *Available at SSRN 4325117*.

Callard, A. (2018). *Aspiration: The agency of becoming*. Oxford University Press.

Chater, N. (2022). What is the point of behavioral public policy? A contractarian approach. *Behavioural Public Policy*, 1-15.

[9] Fabian et al. (2022, 14) describe the idea of coproduction as "a well-established practice in health and design sciences, local governance, and social policy. Coproduction requires an equal and deliberative process that includes and represents the views of relevant stakeholders, with technical experts classed as merely one stakeholder group among many."

Christman, J. (2005). Autonomy, Self-Knowledge, and Liberal Legitimacy. In J. Christman & J. Anderson (Eds.), *Autonomy and the Challenges to Liberalism: New Essays*, 330-358. Cambridge University Press.

Christman, J. (2009). *The politics of persons: Individual autonomy and socio-historical selves*. Cambridge University Press.

Cowen, T. (1993). The scope and limits of preference sovereignty. *Economics & Philosophy*, *9*(2), 253-269.

De Rosa, D., Reggiani, T., & Santori, P. (2021). Special Issue: "The Community of Advantage". *International Review of Economics*, *68*(1), 1-4

Delmotte, C., & Dold, M. (2021). Dynamic preferences and the behavioral case against sin taxes. *Constitutional Political Economy*, 1-20.

Dold, M. (2018). Back to Buchanan? Explorations of welfare and subjectivism in behavioral economics. *Journal of Economic Methodology*, *25*(2), 160-178.

Dold, M., & Lewis, P. (2023). A Neglected Topos in Behavioral Normative Economics: The Opportunity and Process Aspect of Freedom. *Behavioral Public Policy*.

Dold, M., & Rizzo, M. J. (2021). Old Chicago against static welfare economics. *The Journal of Legal Studies*, *50*(2), 179-198.

Dold, M., & Stanton, A. (2021). I choose for myself, therefore I am: The contours of existentialist behavioral economics. *Erasmus Journal for Philosophy and Economics*, 14(1), 1-29.

Dold, M., van Emmerick, E., & Fabian, M. (2022). Taking Psychology Seriously: A Self-Determination Theory Perspective on Sugden's Opportunity Criterion. DOI: 10.13140/RG.2.2.12287.43685/1.

Elster, J. ([1985] 2016). *Sour grapes*. Cambridge University Press.

Fabian, M., & Dold, M. (2022). Agentic preferences: a foundation for nudging when preferences are endogenous. *Behavioural Public Policy*, 1-21. DOI: https://doi.org/10.1017/bpp.2022.17.

Fabian, M., Alexandrova, A., Coyle, D., Agarwala, M., & Felici, M. (2022). Respecting the subject in wellbeing public policy: beyond the social planner perspective. *Journal of European Public Policy*, 1-24.

Frank, R. H. (2021). *Under the Influence*. Princeton University Press.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of economic literature*, *40*(2), 351-401.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

Gigerenzer, G. (forthcoming). From Bounded Rationality to Ecological Rationality. To appear in Gigerenzer, G., Mousavi, S., & Viale, R. (eds.). The Herbert Simon Companion. Elgar Publishing.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*(2), 53–96.

Grüne-Yanoff, T. (2016). Why behavioural policy needs mechanistic evidence. *Economics & Philosophy*, *32*(3), 463-483.

Grüne-Yanoff, T. (2021). Boosts: A remedy for Rizzo and Whitman's Panglossian fatalism. *Review of Behavioral Economics*, *8*(3-4), 285-303.

Grüne-Yanoff, T. (2022). What preferences for behavioral welfare economics? *Journal of Economic Methodology*, *29*(2), 153-165.

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, *26*(1), 149-183.

Grüne-Yanoff, T., Marchionni, C., & Feufel, M. A. (2018). Toward a framework for selecting behavioural policies: How to choose between boosts and nudges. *Economics & Philosophy*, *34*(2), 243-266.

Hargreaves Heap, S. P. (2013). What is the meaning of behavioural economics?. *Cambridge Journal of Economics*, *37*(5), 985-1000.

Hargreaves Heap, S. P. (2017). Behavioral public policy: the constitutional approach. *Behavioral Public Policy*, *1*(2), 252-265.

Hargreaves Heap, S. P. (2020). Structure and agency: themes from experimental economics. In *A research agenda for critical political economy*, 107-120. Edward Elgar Publishing.

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973-986.

Hershfield, H. E., Goldstein, D. G., Sharpe, W. F., Fox, J., Yeykelis, L., Carstensen, L. L., & Bailenson, J. N. (2011). Increasing saving behavior through age-progressed renderings of the future self. *Journal of Marketing Research, 48*, 23–37.

Hoff, K., & Stiglitz, J. E. (2016). Striving for balance in economics: Towards a theory of the social determination of behavior. *Journal of Economic Behavior & Organization*, *126*, 25-57.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes*, *67*(3), 247-257.

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

Krijnen, J. M., Tannenbaum, D., & Fox, C. R. (2017). Choice architecture 2.0: Behavioral policy as an implicit social interaction. *Behavioral Science & Policy*, *3*(2), i-18.

Lichtenstein, S., & Slovic, P. (2006). The construction of preference. Cambridge University Press.

Loewenstein, G., & Angner, E. (2003). Predicting and indulging changing preferences. *Time and decision: Economic and psychological perspectives on intertemporal choice*, 351-91.

Loewenstein, G. F., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological review*, *100*(1), 91-108.

Marewski, J. N., & Gigerenzer, G. (2022). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*(1), 77-89.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. Trends in cognitive sciences, 9(6), 296-305.

Matson, E. W. (2022). Our dynamic being within: Smithian challenges to the new paternalism. *Journal of Economic Methodology, 29*(4), 309-325.

McAdams, D. P. (1996). Personality, modernity, and the storied self: A contemporary framework for studying persons. *Psychological inquiry*, *7*(4), 295-32

McKenzie, C. R., Sher, S., Leong, L. M., & Müller-Trede, J. (2018). Constructed preferences, rationality, and choice architecture. *Review of Behavioral Economics*, *5*(3-4), 337-360.

Milkman, K. L., Minson, J. A., & Volpp, K. G. (2014). Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management science*, *60*(2), 283-299.

Pigou, A. C. (1920). *The Economics of Welfare*. London: Macmillan

Read, D. (2006). Which side are you on? The ethics of self-command. *Journal of Economic Psychology*, *27*(5), 681-693.

Rizzo, M. J., & Whitman, G. (2020). *Escaping paternalism: Rationality, behavioral economics, and public policy*. Cambridge University Press.

Roberts, R. (2022). *Wild Problems: A Guide to the Decisions That Define Us*. Portfolio.

Ryan, R. M., & Deci, E. L. (2006). Self-regulation and the problem of human autonomy: Does psychology need choice, self-determination, and will? *Journal of personality*, *74*(6), 1557-1586.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and Extrinsic Motivation from a Self-Determination Theory Perspective: Definitions, Theory, Practices, and Future Directions. *Contemporary Educational Psychology.*

Schubert, C. (2015). Opportunity and preference learning. *Economics & Philosophy*, *31*(2), 275-295.

Schubert, C. (2021). Opportunity meets self-constitution. *International Review of Economics*, *68*(1), 51-65.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology*, *130*(3), 380-400.

Sheldon, K. M. (2022). *Freely Determined: What the New Psychology of the Self Teaches Us about how to Live.* Basic Books.

Sher, S., McKenzie, C. R., Müller-Trede, J., & Leong, L. (2022). Rational Choice in Context. *Current Directions in Psychological Science, 31*(6), 518-525.

Sen, A. (1997). Maximization and the Act of Choice. *Econometrica, 65*(4), 745-779.

Sen, A. (1999). *Development as Freedom*. Oxford University Press.

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

Sunstein, C. R. (2020). Behavioral welfare economics. *Journal of Benefit-Cost Analysis, 11*(2), 196-220.

Taylor, C. (1989). *Sources of the Self*. Harvard University Press.

Vromen, J., & Aydinonat, N. E. (2021). Introduction to the Review Symposium on Robert Sugden's The Community of Advantage. *Journal of Economic Methodology*, *28*(4), 349-349.