Economics & Philosophy (2026), **42**, 154–173
doi:10.1017/S0266267125100436

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Preferences, goals and implications for paternalism

Petr Krautwurm[1] and Philipp C. Wichardt[2,3,4,5]

[1]Department of Econometrics, Prague University of Economics and Business, Prague, Czech Republic, [2]Department of Economics, Kiel Institute for the World Economy, Kiel, Germany, [3]Department of Economics, University of Lund, Lund, Sweden, [4]Department of Economics, University of Rostock, Rostock, Germany and [5]CESifo Munich, Germany
Corresponding author: Philipp C. Wichardt; Email: philipp.wichardt@uni-rostock.de

(Received 20 December 2024; revised 21 May 2025; accepted 15 June 2025; first published online 11 August 2025)

Abstract
This paper proposes a conceptual model of decision-making tying specific preferences to broader individual goals. Specifically, we consider terminal goals, representing fundamental objectives, and instrumental goals, serving as complexity-reducing intermediate steps toward achieving terminal goals and determining eventual preferences. Notably, the hierarchical goal structure allows for contextual misalignments between different instrumental goals, which may lead to suboptimal decisions – as evaluated from an outside perspective. Thus, applied to the discussion about nudging and paternalism, the model provides a methodological justification for paternalistic interventions as it is compatible with arguments in favour of interventions aimed to correct such choices.

Keywords: Preferences; goals; paternalism; behavioural inconsistencies

1. Introduction

Economic modelling builds on the assumption that agents choose what they prefer. That means, within the model, preferences are an expression of an individual ordering of possible (expected) outcomes and agents are the sovereigns of their own welfare. Moreover, for the theory to have empirical content, it has to be assumed that the underlying preferences are stable, at least to some extent (cf. Becker 1976). Otherwise, no meaningful forecasts or welfare judgements could be made and all behaviours could simply be rationalized as an expression of instantaneous preferences. Little would be gained. Fortunately, substantial evidence suggests that decision-makers' preferences, beliefs and resulting behaviours contain stable and systematic elements (e.g. Amir and Levav 2008; Carlsson *et al.* 2014; O'Grady 2017; Restrepo and Vaisey 2024). However, there is also abundant evidence on the

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

apparent inconsistency of individual choices (e.g. Tversky and Kahneman 1986; Camerer 2003; DellaVigna 2009; Thaler 2015). A challenging natural question arising from this empirical tension is to what extent individuals really are the sovereigns of their own welfare, as suggested by the model.

Notably, questions about decision-maker sovereignty have given rise to much debate about third-party interventions in decision-making in recent years in connection with nudging (cf. Thaler and Sunstein 2008; Hausman and Welch 2009; Thaler 2015; Whitman and Rizzo 2015; Infante *et al.* 2016; Sugden 2017; Sunstein 2018; Kemper and Wichardt 2024a). Introduced in 2008, nudging builds on the observation that the decisions people make tend to be influenced by seemingly minor details of their decision environment. Observing this and arguing that such deviations may be harmful to the individual, nudging intends "[to] influence choices in a way that will make choosers better off *as judged by themselves*" (Thaler and Sunstein 2008: 5, italics in original) by appropriately changing such details without altering available options.

Since its introduction, the idea has gained considerable attention not only in academic but also in political circles (see Leggett 2014, for a discussion and references) as it suggests that tangible changes are available through seemingly innocuous means. The part of the idea that has given rise to much debate is the one expressed in italics – "better off, *as judged by themselves*" – as it presumes that it is possible to assess from the outside what decision-makers would have wanted despite not choosing it when they could have. Put simply, if decision-makers are perfect sovereigns of their welfare, they always make the best possible choices given their options, rendering third-party interventions inefficient. If decision-makers are prone to errors, third-party interventions may indeed improve their chances of making better decisions according to their own assessment. Yet, even if we allow for the theoretical possibility of errors, how can we judge which choices are based on 'true' preferences and which are erroneous within the common model of preferences (cf. Hausman and Welch 2009; Whitman and Rizzo 2015; Infante *et al.* 2016; Häußermann 2019; Špecián 2019; Kemper and Wichardt 2024a; Colin-Jaeger and Dold 2025; Fabian and Dold 2025)?

Arguing along these lines Špecián (2019), for example, warns of a paradox inherent in paternalistic thinking, in which paternalists must first prove their ability to identify other preferences than those agents currently follow while ensuring that these preferences are the authentic, true preferences that individuals would ideally want to have but fail to do so. Similarly, Kemper and Wichardt (2024a) highlight an analogous issue in the definition of welfare, which must inherently be tied to subjective aspects of preferences. As long as choices are taken as a sovereign expression of the agent's preferences and all preferences are on equal footing, this problem is difficult to overcome as, within the standard model framework, there is no way of prioritizing one set of preferences inferred from choices over another. Accordingly, the standard model itself provides no reason to argue for one choice being a proper one and another one being an an error requiring correction through interference.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

Yet, proponents of such external interventions, particularly nudging, keep emphasizing that nudging is not meant to be paternalistic (e.g. Thaler 2015).[1] In fact, Thaler writes "A point that critics of our book seem incapable of getting: we have no interest in telling people what to do" and immediately continues "We want to help them achieve their *own* goals" (Thaler 2015: 325, italics in original). But even if we assume that it may be practically possible to help people in this way, how can we judge what people want based on a sound theoretical framework?

In the present paper, we propose a modification of the common model of preferences which leaves room for external judgements about individual benefits. In doing so, we start from the observation that (part of) the tension in the debate arises from the fact that standard arguments about preferences offer no inherent hierarchy. Moreover, we note that Thaler in the quote cited above does not mention preferences at all but instead refers to goals (see also Sen 1985). In fact, goals and goal-directed behaviour are standard terms in the psychological literature (see, for instance, Deci and Ryan 2000 or Dold *et al.* 2024, for prominent examples and additional references). Notably, the distinction between goals and preferences, we believe, offers a possible way out of (at least part of) the dilemma as goals are easy to imagine being hierarchically ordered.

The proposed conceptual model of preferences is built on the notion that decision-makers follow different types of goals. For the sake of argument, we consider two types of goals, namely terminal and instrumental goals. By assumption, terminal goals represent more fundamental objectives like living a long and healthy life, while instrumental goals correspond to more intermediate objectives directed towards reaching terminal goals, like eating healthily. Individual preferences, then, are derived from the instrumental goals relevant in a certain context (regarding contextual effects on behaviour, see also Bergh and Wichardt 2018; Dold 2018; Delmotte and Dold 2022; Kemper and Wichardt 2024b). Thus, intuitively, instrumental goals can be understood as a complexity-reducing mediator between terminal goals and daily-life decision-making. What is important to note is that the conceptual move to consider hierarchical ordering of goals now allows for both: stable behavioural patterns directed towards terminal goals and context dependent deviations from these patterns (which can still be best responses to circumstances).

In order to exemplify this point, consider an agent, say Alfons, with terminal goals 'social connectedness' and 'long and healthy life'. Moreover, assume that Alfons's corresponding instrumental goals are *education*, *social activities*, *healthy eating* and *exercising*. In such a situation, it seems reasonable to expect that a long-term observation of Alfons would show comparably stable patterns towards education, meeting people, eating healthily and exercising regularly. Yet, if observed in a specific context with friends who favour fast food, it still seems reasonable to assume that Alfons would join them in their choice of food in order not to stand out

[1]While we do not want to enter into semantic debates here, it is worth noting that according to Conly (2012) definition of paternalism, nudging would, in any case, be paternalistic. This, however, still does not answer the more central questions whether it is so in a way that could reasonably be argued for as problematic. Additionally, the appropriate definition of paternalism is not a topic of the present paper.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

(considering that a single deviation from healthy eating could be relatively inconsequential).

Similarly, taking up the much quoted cafeteria example of Thaler and Sunstein (2008), urgent contextual needs such as hunger paired with inattention may well lead Alfons to pick the cake in the cafeteria instead of the salad, if the latter is placed less prominently. Yet, neither choice would have to be considered a mistake in the respective instance as it might be the best Alfons can do given the circumstances. However, in both cases, third-party intervention could arguably help Alfons make better choices *as judged by himself*. In the first case, for example, publicly promoting healthy eating and encouraging acceptance of those who choose to eat more healthily might weaken the pressure on Alfons to copy his friends' behaviour or even motivate his friends to change.[2] In the latter case, swapping the positions of salad and cake in the cafeteria, as suggested by Thaler and Sunstein (2008), might already be enough to encourage Alfons to choose the salad, which could be argued for as being in line with his (terminal) goals.

The simple but crucial point to note is that the hierarchy of goals allows for judging actual choices as, in fact, the best possible in the specific situation (referring to preferences derived from instrumental goals) as well as providing a basis for arguing that third-party interventions may be beneficial from Alfons's broader perspective (in view of corresponding terminal goals). The room for improvement comes from accepting the agent's bounded rationality in the process of creating actual preferences. As terminal goals are inherently complex to follow, agents decompose these complex tasks into simpler parts by creating instrumental goals, which serve as intermediaries. These instrumental goals can be more easily evaluated by agents once put into a specific decision context. This way, an agent can implicitly acknowledge the broader goal of a long and healthy life by following instrumental goals, such as eating healthily and increasing vegetable consumption (see also Fabian and Dold 2025 for a similar example). Figure 1 illustrates how a more complete ordering may look like. This, however, may lead to occasional misalignments due to discrepancies between what seems best *now* and what would be better from a broader perspective.

Thus, the proposed solution to the paternalistic paradox identified by Špecián (2019) lies in recognizing that decision-makers may have terminal goals which are inherently complex to follow, making it nearly impossible for agents to determine how individual actions affect their terminal goals. Consequently, agents decompose these complex tasks into simpler parts by creating instrumental goals, which serve as intermediaries. This way, it becomes easier for agents to evaluate specific choices by judging how different options would satisfy these intermediate goals, thereby still helping them to achieve their terminal goals, while accepting some suboptimal choices as a price for complexity reduction. For example, to survive, agents may recognize the need for eating healthily, leading to the intermediate goal of increasing vegetable consumption, as depicted in Figure 1. Eating a salad contributes to the instrumental goal of increasing vegetable intake, which in turn contributes to eating healthily, ultimately contributing to survival. Yet, as we have seen above, contextual

[2]Note that external intervention in this case is difficult as complying with the group is not a mistake per se but does serve another terminal goal.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

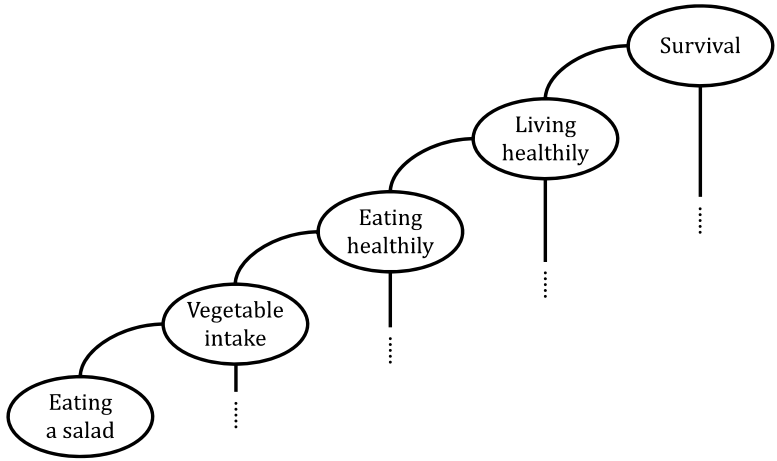


Figure 1. Contribution of eating a salad to survival. While we consider only two levels in our model, there can be more and each of the higher goals can have multiple goals that contribute to them (as illustrated here; additional lines indicating possible further avenues).

aspects may lead the agent to deviate from pursuing certain goals (locally) – e.g. because of other instrumental goals being deemed more relevant in the moment (i.e. a misalignment of goals). In these cases, there can be room for paternalists – parents, friends, policymakers, etc. – reasonably justifying interventions in the agents' decision environment (as put by Thaler 2015: 325; italics in original) "to help them achieve their *own* goals".

While we are confident that the proposed perspective on preferences mitigates the paradox described by Špecián (2019), it should be clear that it does not solve all the problems there are. First of all, the argument requires a sound and reliable analysis of long-run patterns in behaviour to infer what we refer to as terminal goals. This in itself is not without the risk of mistakes, and failure to do so properly could result in paternalists harming the very individuals they aim to help. Moreover, decision-makers need not recognize that they are deviating from their terminal goals so that interventions could trigger psychological reactance (e.g. Rains 2012), causing agents to resent the perceived intrusion on their autonomy.[3] Finally, different decision-makers may prioritize their terminal goals differently, requiring tailored types of interventions, as flat, one-size-fits-all interventions are unlikely to benefit everyone. Given these challenges, we conclude that, while efficient paternalism is theoretically possible, it is fraught with practical difficulties that make it challenging to implement successfully (i.e. to the benefit of the agents). Thus, even if seen from the perspective presented in this paper, many interventions will still not be as innocuous as suggested by Thaler and Sunstein (2008).

The rest of the paper is structured as follows: in section 2, we introduce the conceptual model and present a more detailed argument supporting its key

[3]We assume that as agents aim to achieve terminal goals through their instrumental goals, they would modify the instrumental goals if they thought they were misaligned with the corresponding terminal goals.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

components as well as general behavioural implications. Section 3 considers possible applications of the argument, both specific to the nudging debate (section 3.1) and to broader paternalistic interventions (section 3.2). Section 4 concludes.

2. Preferences, Goals and Behaviour

This section is split into two parts: a motivation of the conceptual model (section 2.1) and a discussion of general behavioural implications (section 2.2).

2.1 The Conceptual Model

As a first step, consider a decision-maker whose behaviour can be conceptualized as following a distinct hierarchy (of sets) of goals from which preferences are derived.

2.1.1 Basic terminology

To avoid definitional issues, we follow the theory of revealed preferences (Samuelson 1938, 1948) in that we assume that decision-makers choose what they prefer.[4] However, these preferences are tied to the moment and the context of decision-making. Thus, when decision-makers choose one action over another in some context, this does not imply that they would do so across different contexts. Accordingly, by definition, there is no way of reliably inferring future preferences from current behaviour. If an agent makes a particular choice, this choice is considered to be preferred over all other options available at that specific moment – not more and not less. This means that, within the model, it is impossible for decision-makers to act against their preferences at the time of decision-making.

Unlike previous conceptions, we do not take these (contextual) preferences as fundamental but as derived from the agent's goals, where goals can be conceived of as answers to the broader, context-independent question of what the agent wants to achieve in their life. Examples would be social connection, a long and healthy life, joining a sports club, or obtaining a more interesting job (where the first two examples refer to more general long-term goals while the latter are more specific and short term). Yet, in contrast to preferences, we assume that goals are hierarchically ordered reflecting the importance the agents assign to them in their life; cf. Figure 2.

2.1.2 Terminal and instrumental goals

For the sake of argument, we restrict attention to a two-level hierarchy and assume that goals can be categorized as either terminal or instrumental (cf. Brandtstädter and Lerner 1999; Bostrom 2012; Ford and Ford 2019); we consider only one level of instrumental goals since additional levels are redundant for the purposes of our analysis. In this terminology, terminal goals represent fundamental, long-term objectives, such as living a long and healthy life, while instrumental goals

[4]For the sake of clarity, we exclude reflexive behaviours – those actions that occur automatically and without conscious intention – from our primary discussion.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

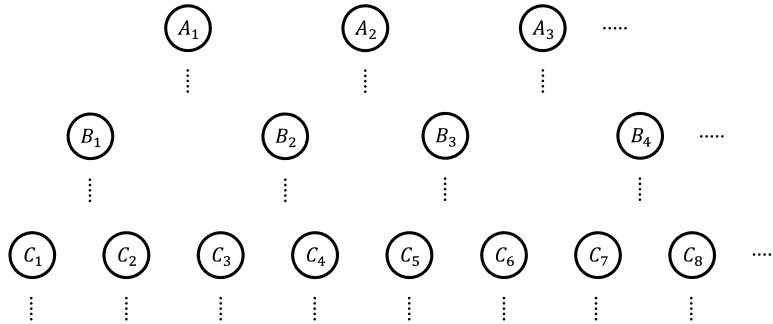


Figure 2. Illustration of a goal hierarchy; different letters indicating different levels of the hierarchy and lower levels being more specific (in the model, reduced to two levels).

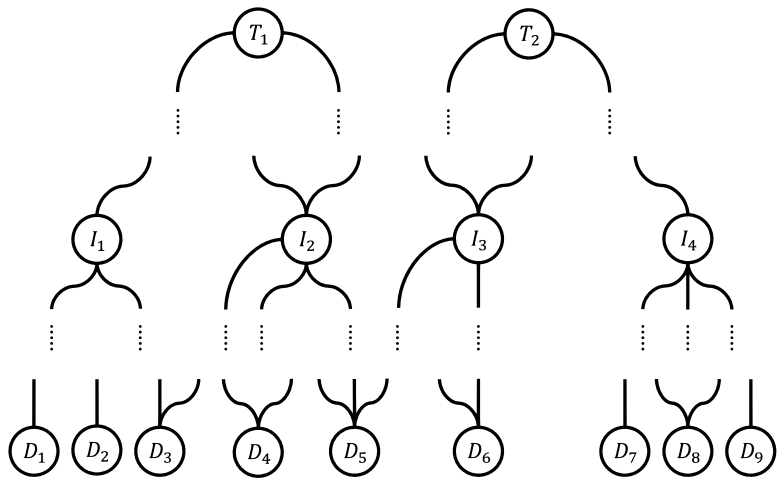


Figure 3. Illustration of goal hierarchy, focusing on the connection between terminal goals and a decision level mediated by intermediate goals in between (in the model, reduced to two levels). Letters signify the level of goals (T for terminal, I for intermediate) and (D) reflecting decision relevant preferences, respectively.

correspond to intermediate, more specific steps, such as eating healthily or exercising regularly guiding actual behaviour; cf. Figure 3.

Intuitively, the hierarchical structure reflects the fact that the complexity of life in combination with the agents' limitations (e.g. Cowan 2010) renders it impossible for real-life agents to always perfectly assess the contribution of all available actions to their different fundamental goals, an observation that is compatible with abundant empirical evidence (e.g. Camerer 2003; Thaler 2015). In that sense, the present model can be conceived of as a model of bounded rationality. Thus, instrumental goals provide cognitive shortcuts for otherwise rational decision-makers, breaking down complex decisions into smaller, manageable tasks. Finally, actual preferences are tied to the lowest level of goals in the hierarchy (see also Fabian and Dold 2025).

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

To exemplify this point, consider an agent with a terminal goal of a long and healthy life. The overarching goal of maintaining good health, for example, is not just about eating vegetables, but encompasses a wide range of behaviours, including regular but not excessive exercise (e.g. Meyer *et al.* 2011; Ruegsegger and Booth 2017), a balanced diet emphasizing food synergy (e.g. Jacobs and Tapsell 2013), appropriate legume intake (e.g. Polak *et al.* 2015), and moderate meat consumption (e.g. Biesalski 2005), along with stress management (e.g. McEwen 2008) and finding meaning in life (Hooker *et al.* 2018). Evaluating how specific actions contribute to this broad objective in every single context is arguably complex, even if we neglect issues arising from potentially conflicting terminal goals (see section 2.2 for a discussion).

Take, for instance, the act of eating a salad: although there is likely a connection between eating healthily and increased survivability (Polak *et al.* 2015), without deeper analysis, the precise nature of this connection is difficult to assess for the average agent even in terms of expected outcomes, let alone in every possible context.[5] Given the difficulty of evaluating the impact of single actions on terminal goals, agents benefit from choosing more manageable goals that have clearer links to higher-level objectives. Consequently, a rational response of cognitively limited agents is to create instrumental goals. By breaking down high-level goals into specific, measurable tasks, agents can more easily monitor their progress and make efficient choices without overwhelming their cognitive capacities. Thus, by following the (lowest level of) instrumental goals, agents respond rationally to a complex environment while acknowledging their own constraints.

2.1.3 About goals

Regarding the division of goals into two levels, it should be noted that the focus on two is chosen simply for the ease of exposition of the main argument. As indicated in the introductory example (cf. Figure 1), a more differentiated structure would be easy to motivate. The basic argument, however, would remain the same. What is important for the subsequent discussion, though, is that agents may have multiple terminal goals and that goals and their relative importance for decision-makers are likely to be idiosyncratic. In that sense, goals can, for example, be viewed as linked to personality, which describes how individuals generally act, while goals explain why they act as they do.

Considering the suggested relationship between personality and terminal goals also squares well with the idea of multiple terminal goals. For example, evolutionary theorists suggest that individual differences within species are linked to variations in personality traits (e.g. Nettle 2006; Cote *et al.* 2008; Gosling 2008). In this context, the framework of the Big Five irreducible personality traits (e.g. Roberts and Robins 2000; McAdams and Pals 2006; Raggatt 2006) provides a compelling argument for the existence of multiple terminal goals. Similarly, studies on human goal structures

[5]Note that genetic predispositions, current health status, environmental conditions, etc. may all be relevant for such assessments. Also, two seemingly very similar actions, such as choosing between salmon or tuna for dinner, may have distinct implications for overarching goals, due to slight differences in their nutritional and contaminant profile (Shim *et al.* 2004), but these differences might not be obvious without experimentation or prior research.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

show that individuals typically have between three and five significant terminal goals and further substantiate the existence of a hierarchical organization of goals (e.g. Chulef *et al.* 2001; Talevich *et al.* 2017). Furthermore, Rokeach (1973) identified eighteen terminal values that people may hold, which is relevant to our discussion, as values can also be interpreted as goals (e.g. Schwartz 1992). In either case, terminal goals can be understood as representing fundamental, desirable aspects of life.

To provide some specific examples, goals which could reasonably be grouped as terminal would be truth-seeking (e.g. Grimm 2008), fairness (e.g. Loewenstein *et al.* 1989; Tabibnia *et al.* 2008; McAvoy *et al.* 2022), procreation (e.g. Bühler 1964), survival (e.g. Kaplan and Gangestad 2015), social prestige (e.g. Zakharenko 2016), and social relatedness (e.g. Hicks and King 2009).[6] Once again, what is important for the present discussion is not the exact specification or grouping of goals but their hierarchical structure (and that the cardinality of terminal goals is larger than one).

2.1.4 Goal hierarchy, context and decision-making

Finally, there are some additional aspects of our model which require brief commenting.

First of all, we (implicitly) assume that the number of goals increases with each level further down in the goal hierarchy. The necessity for this follows immediately from the underlying idea of complexity reduction. If the number of goals remained constant, agents would merely exchange one set of goals for another. If the number of goals was smaller on lower levels than on higher ones, the information embedded in those goals would increase rather than decrease. Accordingly, for compatibility with the underlying intuition, the number of goals has to increase with each level further down in the hierarchy.

Moreover, once a specific choice in a specific context is considered, agents may be faced with the problem of trade-offs between multiple goals at the same level. They can solve this, for example, by weighting goals (and neglecting those which are not relevant for the decision at hand, i.e. assigning zero importance to them in the weighing). These trade-offs between relevant goals apply across all levels of the goal hierarchy and may lead to different outcomes for different decision-makers. For instance, Socrates sacrificed his own survival by refusing to escape his sentence in order to convey a message about upholding what is right (cf. Barker 1977). War heroes often prioritize camaraderie and collective survival over their own lives (cf. Atran *et al.* 2014). Others may trade their lives for an idea or to secure the financial stability of their family, as seen in cases where individuals time their deaths to maximize inheritance (cf. Kopczuk and Slemrod 2003). Contrary to that, some people may prioritize personal pleasures over social prestige (cf. Baumeister and Scher 1988). Therefore, different agents can prioritize distinct goals, resulting in

[6]From an evolutionary perspective, one may argue that all human behaviour is ultimately aimed at maximizing inclusive fitness (e.g. Hamilton 1964), with all other goals acting as instrumental to this overarching objective. We do not follow this view, proposing that inclusive fitness is an emergent property. In our view, considering multiple terminal goals provides a more useful framework for understanding diverse preferences.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

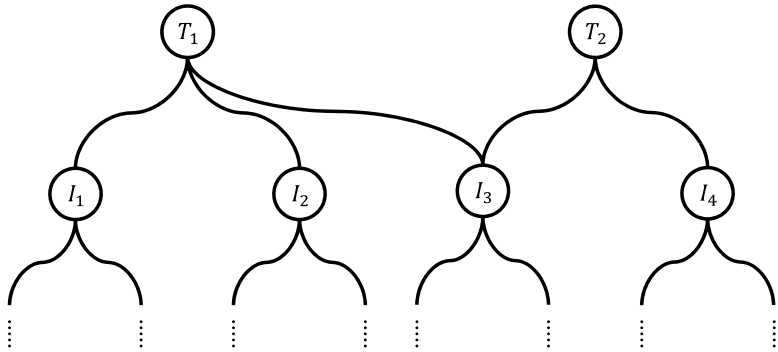


Figure 4. Illustration of a case where one instrumental goal serves two different terminal goals. As in the example, T_1 could be social connection, T_2 health and I_3 joining team sports.

different trade-offs. Moreover, even for a single agent, what is driving behaviour in one context need not be the same in another context.

Yet, while behaviour may seem inconsistent between two different contexts, the model would suggest stable general patterns in behaviour directed towards terminal goals. Thus, observing that behaviour is commonly found not to be entirely random but exhibiting some degree of stability (Amir and Levav 2008; Carlsson *et al.* 2014; O'Grady 2017; Restrepo and Vaisey 2024), the model would provide a conceptual rationale for such recurring patterns as well as for the apparent (local) instability of behaviour.

To wit, within the model decision-makers have two levels of goals and context-specific preferences over available outcomes derive from the lowest level of goals (appropriately weighting relevant goals); see Figure 3 for illustration.

2.2 Behavioural Aspects

Equipped with the conceptual idea and some motivation, we move on to explore the broader implications of the proposed model for behaviour. In particular, we discuss the connection between contextual preferences and different, possibly conflicting goals.

2.2.1 Alignment and misalignments of goals

First of all, it is important to recognize that instrumental goals can serve more than one terminal goal. For example, engaging in team sports would serve both social connection and a long and healthy life; see Figure 4 for illustration. However, instrumental goals derived from different terminal goals may also come into conflicts leading to decisions being in line with one terminal goal but not the other – such as in the introductory example of eating a salad (healthy eating and long and healthy life) while being with friends who are ordering fast food (social connection).[7] Moreover, depending on circumstances, instrumental goals may

[7]Here, we abstract from the fact that also social connection is relevant for good health.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

even induce contextual preferences for behaviour with different short-term and long-term effects for corresponding terminal goals. For instance, a specific group of agents focused on maintaining good health may develop routines that benefit one aspect of health while harming another – such as drinking alcohol in the evening to relieve stress (e.g. Sillaber and Henniger 2004), which may aid short-term stress reduction but damage internal organs in the long run.

To manage such trade-offs, we again assume that for each choice agents put relative weights on their instrumental goals according to contextual stimuli (for a discussion of contextual effects on decision-making, see also Bergh and Wichardt 2018; Dold *et al.* 2024; Kemper and Wichardt 2024b; Colin-Jaeger and Dold 2025; Fabian and Dold 2025). Thus, in the drinking example, agents may conclude that in view of the current situation stress relief outweighs concerns about visceral health. Accordingly, drinking alcohol would contribute more to local optimization (achieving immediate stress relief) than it does to harming long-term health (liver damage, etc.), possibly even leading to habit formation through developing physical dependence. Similarly, in the eating with friends example, agents may prioritize not offending their friends over eating healthily as it is (locally) judged to be more rewarding.[8]

What is important for the present argument is that instrumental goals may cause such conflicts to emerge and that behaviour, therefore, need not always be aligned with terminal goals and that such non-alginments are not necessarily a sign of irrationality.

2.2.2 Dynamic misalignments

A further source of structural inconsistencies, which deserves a brief mention, stems from dynamic changes in the agent's life; for example, resulting from significant life events that may prompt agents to reassess their priorities (cf. Ojanen *et al.* 2007). In fact, research on the Big Five personality traits indicates that these can evolve over time, often following identifiable trends (cf. Mroczek and Spiro 2003; Roberts and Mroczek 2008). Additionally, Conway and Williams (2008) emphasize that shifts in how individuals perceive themselves are closely linked to changes in their goals. Moreover, Bühler (1964) provides direct evidence of goals evolving over time, often in discernible patterns.[9] Thus, as terminal goals evolve, even small changes in their weighting can necessitate substantial adjustments throughout the goal hierarchy.[10] Recall that instrumental goals are thought of as complexity-reducing intermediaries

[8]Here, peer pressure (e.g. Bonein and Denant-Boèmont 2015) would be interpreted as increasing the salience of an instrumental goal such as "not offending friends" derived from "social connection".

[9]This is supported by Brandtstädter and Lerner (1999) and Habermas and Bluck (2000), whose findings also suggest that deviations are more common at the instrumental level than at the level of terminal goals and that misalignments should diminish over time. However, both partial and overarching goals in life can change (King and Hicks 2007; Grahek *et al.* 2023).

[10]Burke (2006) identifies two sources of changes in how individuals perceive themselves over time: one is gradual and tied to an individual's general self-view, while the other is more volatile and linked to conflicts in the meaning of specific life aspects. The former arguably can be seen as corresponding to changes in the weights of terminal goals, while the latter may indicate misalignments between different instrumental goals. This suggests that changes in the weights of terminal goals occur more slowly than changes at lower levels of the goal hierarchy.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

to achieve terminal goals. Thus, if a change in the agent's circumstances induces adjustments in (the relative importance of) terminal goals, this may imply temporary misalignments between instrumental and terminal goals, as well as an adjustment process.

While we do not explicitly consider learning, it is important to acknowledge the possibility of such changes and the resulting need for adaptation. In fact, Bostrom (2012) interprets misalignments between terminal and instrumental goals as a sign of irrationality. However, this is not the only possible interpretation as, for example, a mere lack of self-knowledge (Tirole 2002) may contribute to these misalignments. Additionally, decision-makers may be uncertain about how certain instrumental goals serve their terminal goals and may be in a phase of exploring alternatives (Cohen *et al.* 2007), something Bostrom (2012) also acknowledges; Fabian and Dold (2025) consider the case of agents learning their preferences, which can be interpreted as understanding which instrumental goals support which terminal goals. Moreover, adjustments are not without cost, as they require time and cognitive resources to observe, evaluate and establish new goals (Grahek *et al.* 2023). As a result, goal adjustments are not instantaneous. Once again, what is important here is that there are perfectly rational reasons for many temporary misalignments and seeming inconsistencies in behaviour.

2.2.3 Resulting patterns

The preceding observations suggest that while decision-makers use goal structures to deduce preferences and to simplify their decision-making processes, these structures inherently produce misalignments between different goals. Depending on the context of a decision, these may occasionally lead to decisions which, upon more deliberate reflection, may not be in the interest of the agent's broader terminal goals. Put differently, the proposed conceptual model of goal hierarchies and preferences is compatible with some degree of seeming behavioural 'inconsistencies' showing in single instances as well as with more consistent overall behavioural patterns. As we argue below, it is the possibility of seeming inconsistencies that provides grounds for justifications for paternalistic interventions.

3. Discussion

In this section, we discuss implications of the proposed conceptual model for policy topics related to paternalistic interventions of various sorts. The purpose of this discussion is to illustrate why a more detailed conception of preferences as derived from goals is helpful to mitigate some of the tension in the debate. As a first step, we take up the nudging debate (3.1), before moving on to paternalistic interventions more generally (3.2), and addressing remaining practical challenges (3.3).

3.1 Nudging

First, we consider the claims made by Thaler and Sunstein (2008) in connection with nudging, in particular, their argument that influencing the behaviour of others can enhance their welfare as judged by themselves.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

3.1.1 Nudging – better off as judged by themselves

As already emphasized in the Introduction, the crucial claim made by Thaler and Sunstein (2008) in connection with their proposal to change decision environments is that such interventions are thought "[to] influence choices in a way that will make choosers better off *as judged by themselves*" (Thaler and Sunstein 2008: 5; italics in original). It is this claim which has fuelled much of the critical debate around nudging (e.g. Hausman and Welch 2009; Whitman and Rizzo 2015; Infante *et al.* 2016; Sugden 2017; Kemper and Wichardt 2024a). As a central aspect of nudging is to not change available options, the debate essentially revolves around the question on what grounds an external person can determine what would have been better for an agent to do according to their own judgement despite the fact that they did not choose to do so.

Note that in order to answer such a question, two sets of preferences have to be considered: one that reflects the behaviour agents currently exhibit, and another representing how they would have preferred to behave but did not. It is in this context that Špecián (2019) highlights a seemingly contradictory challenge for libertarian paternalists: they must identify normatively binding preferences that differ from the current (revealed) preferences of the agents, while ensuring that these preferences are still derived from relevant aspects of the agents' preferences.[11] If, within the model, preferences are assumed to be stable, this challenge is difficult to master. Even if we think of agents as possessing two selves, one for actual decisions and one with an eye on broader developments but each with its own notion of well-being (cf. Thaler and Shefrin 1981), the difficulties would merely be shifted to a different level. Eventually, both would motivate preferences from their own perspective and would do so for good reason. Hence, any internal compromise reached for a specific decision would be difficult for external parties to challenge without questioning the agent's autonomy.

However, once preferences are conceptualized as derived from hierarchically ordered goals, which provide rough guidance for behaviour, it is possible to overcome these methodological difficulties and to rationalize both: a high degree of consistency in individual behaviour (being constantly directed towards certain agent-specific terminal goals) as well as occasional contextual inconsistencies (stemming from contextual misalignments of different instrumental goals). In fact, according to the argument presented in section 2, behaviour can exhibit clearly discernible patterns if observed over a longer period and across different contexts, while locally appearing much more erratic. Moreover, as we assume that terminal goals represent what agents ultimately seek to achieve, relying on the general patterns to interfere with local behaviour (e.g. to nudge agents) would not impose external values. Thus, inasmuch as paternalists would refer to such general patterns in combination with a model of preferences along the lines outlined in the previous section, they may indeed have reason to argue that they intend to improve the agents' welfare as judged by themselves.

In that sense, the proposed model of preferences bridges Thaler and Sunstein (2008) perspective with the theory of revealed preferences (Samuelson 1938, 1948),

[11]Note that what sets paternalists apart from perfectionists is that they – different from perfectionists – cannot simply replace someone's preferences with their own ideal values (Conly 2012).

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

as the normatively binding preferences are revealed as a latent trend around which the observed preferences revolve. Over time, these normatively binding preferences are revealed through the choices agents make, i.e. they are embedded within the patterns of observed behaviour. Thus, the long-run trend gives reason to argue for local interventions – meddling with contextual preferences – being consistent with the agent's broader goals (as Thaler 2015 argues).

3.1.2 Broader implications for paternalistic interventions

Before moving on, it is worth noting that the above argument is not restricted to nudging or libertarian paternalism (as nudging is often referred to). While nudging aims to intervene without changing available options, which might sound particularly innocuous, the central problem is essentially the same for all paternalists in general. What is eventually intended is a change in observable behaviour of an agent and the argument to support such interventions is that they would improve the situation of the acting agent according to their own values. Accordingly, as long as a model of preferences allows one to argue that some immediate behaviour is not in line with the agent's broader goals and that this deviation is worth correcting, it offers a possible justification for interventions. Indeed, the strength of this justification would depend on the plausibility of the model and the reliability of various steps of the argument (cf. section 3.3).

3.2 Beyond Nudging

A specific form of paternalistic interventions, which deserves a brief mention here, relates to policies designed to address changes in people's lives that are arguably not sufficiently considered by decision-makers on their own. Examples include mandatory or socially incentivized retirement savings (e.g. Thaler and Benartzi 2004; Ashraf *et al.* 2006; Gugerty 2007; Kast *et al.* 2018), obligatory health insurance (e.g. Twigg 1999; Erlangga *et al.* 2019; Durizzo *et al.* 2022), or generally incentivized healthy lifestyle (e.g. Gruber and Mullainathan 2005; Giné *et al.* 2010; Schwartz *et al.* 2014). In such cases, a possible argument against these measures is that decisions should be left to individuals, as they are presumed to know what is best for themselves.[12] Such an argument, however, would rely on the traditional conception of preferences, which would not allow for what we might refer to as locally rational mistakes.

Within the present model, a possible rationale in favour of such interventions would involve changes in terminal goals that may occur over time. While we do not consider such changes or learning within the model, it seems reasonable to assume that terminal goals such as *health*, *old age finances*, or *health of children* either change in their relative importance with age or are not considered at all, as long as they are not relevant (see e.g. Austad 1997; Kaplan 1997). Such changes in terminal goals, however, would necessitate adjustments within the goal hierarchy which are likely to take time. In view of this, it may be reasonable to help people adapt to the

[12]An additional argument against such measures is offered by Kemper and Wichardt (2024a) and Fabian and Dold (2025), who suggest that third parties may eventually implement something that is not in the interest of the person being nudged but serves their own purposes.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

new circumstances. For example, policies could promote information campaigns about preventive health examinations (e.g. Suk 2011), including vouchers as a libertarian intervention, or implement obligatory health insurance (e.g. Durizzo *et al.* 2022) or incentivized pension savings (e.g. Gugerty 2007).

Note that the argument would once again be based on the structural conception of preferences proposed in the present paper. The only difference to the argument provided around nudging is that the focus here is not on contextual misalignments for single agents but on general life-cycle developments common to most people. In essence, however, it is again the fact that the hierarchical structure of the model allows for prioritizing one type of judgement over another, a feature that is absent in the common model of preferences and drives our argument.

3.3 Remaining Practical Issues

Finally, we want to emphasize that all paternalistic interventions face practical problems related to the assessment of terminal goals and establishing appropriate restrictions to guide agents toward achieving those goals.

For example, a plausible strategy for paternalists would be to observe past behaviour of decision-makers and to infer a latent trend. Yet, even assuming that terminal goals do not change over time, estimating a trend from any set of data is rarely free of possible measurement and estimation errors. Consequently, paternalists are still methodologically limited in their ability to correctly assess terminal goals, with the obvious consequences for welfare judgements. What is more, even assuming terminal goals to be correctly established, policymakers still have to tailor interventions to suit individual needs. Failure to do so obviously risks creating interventions that successfully alter behaviour but fail to efficiently move agents closer to their terminal goals (cf. Špecián 2022).

Note that also eliciting agents' preferences – after some reflection – in a neutral state where no interference of the decision context is possible (e.g. Beshears *et al.* 2008; Allcott and Sunstein 2015), is not perfectly reliable if terminal goals are to be established.[13] Consider, for example, agents who regularly eat cake in a cafeteria. While the agents may indeed have a long and healthy life as a terminal goal, they may also experience a temporary need for the stress relief provided by the indulgence – which in turn may be good for their overall health due to the local stress relief. In line with the earlier example of drinkers (cf. Sillaber and Henniger 2004), these agents may rationally consider the short-term benefits of stress relief to outweigh the long-term harm to health (caused by sugar and saturated fats). Yet, even if in a neutral state, agents may still present a socially acceptable narrative (or even a narrative they like themselves) to gain social credit (Simler and Hanson 2017). Therefore, we would argue that, as different instrumental goals are relevant in different contexts, also a neutral state is just one specific context (a very unnatural

[13]Deliberative forums may be a reasonable way to find a practical solution in a democratic environment (see Button 2018; Häußermann 2019; Špecián 2022 for a discussion). Nevertheless, they do not dissolve the problem of eliciting agents' underlying long-term goals as any momentary responses are likely to be affected by local circumstances. Eventually, someone external would have to decide which individual statements to go by (see also Kemper and Wichardt 2024a).

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

one). Instead, we believe that taking the average of observed preferences across multiple contexts provides a more balanced and accurate reflection of agents' terminal goals, as it captures the diversity of circumstances in which decisions are made.

Last but not least, since agents themselves may not be able to recognize the long-term ramifications of their short-term choices or potential improvements, they may still perceive external interventions as harmful per se; for example, due to psychological reactance (cf. Rains 2012; Schütze *et al.* 2025). Since agents in the moment of the decision, by definition, believe that they are making the best choice available, the benefit of any third-party intervention can only be acknowledged *ex post*. This feature is intrinsic to the goal hierarchy.

To wit, it is important to note that both terminal goals and appropriate measures needed to help agents achieve them are likely to differ among decision-makers, making it nearly impossible to design a flat intervention that suits everyone. Thus, while single interventions may be good for agents on average (or the majority of agents, etc.), they are unlikely to benefit all individuals uniformly and may even harm some. Accordingly, even if we knew what would be good for each individual, constructing an uncontentious social intervention would still be problematic.

4. Concluding Remarks

In this paper, we have proposed a conceptual model of preferences centred on the interaction between terminal and instrumental goals. As we have argued, the model offers a structured framework to rationalize (some) seeming inconsistencies in decision-making and to support arguments in favour of paternalistic interventions.

More specifically, the model utilizes a distinction between terminal goals, which correspond to fundamental life-defining objectives, and instrumental goals, which serve as intermediate steps to achieve these overarching aims. Intuitively, instrumental goals are thought of as a complexity-reducing mediator between terminal goals and context-specific preferences. In that sense, the model integrates human-bounded rationality into a conceptual model of preferences. As we have argued, this allows for contextual misalignments and "errors" which may be worth correcting – even from the perspective of the agent themselves. Thus, within the model, there is a way to prioritize one perspective on decisions over another because terminal goals are what guides instrumental goals. Accordingly, the model offers a methodological basis to argue for paternalistic interventions as, within the model, behaviour aligned with terminal goals provides a normatively binding benchmark that paternalists can adhere to.

While the proposed conceptual model of preferences offers a theoretical justification for paternalistic interventions, whether in terms of nudges or hard interventions, several practical challenges remain. For one thing, interventions would have to rely on estimations of long-term trends in behaviour to identify terminal goals – with the common uncertainties inherent in any data-based estimation. Also, preferences and goals will differ between individuals and, hence, "one-size-fits-all" policy measures will still be difficult to argue for. Thus, also arguments based on goals and preferences as proposed in this paper would not grant

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

paternalists all the liberty they might desire. Yet, conceiving of preferences as being derived from a hierarchy of goals *does* give a methodological foundation for arguments in favour of interventions "mak[ing] choosers better off *as judged by themselves*" (Thaler and Sunstein 2008: 5; italics in original).

Acknowledgements. We are grateful to Jan Krause, Ulrich Schmidt, an anonymous reviewer and a seminar audience at the Kiel Institute for the World Economy for helpful comments and discussions. Wichardt thanks the Arne Rydes Stiftelse for financial support. Krautwurm thanks the Kiel Institute for the World Economy for its hospitality and The Internal Grant Agency of Prague University of Economics and Business for supporting his work [VSE IGS F4/39/2025]. The usual disclaimer applies.

References

Allcott H. and C.R. Sunstein 2015. Regulating internalities. *Journal of Policy Analysis and Management* **34**, 698–705.

Amir O. and J. Levav 2008. Choice construction versus preference construction: the instability of preferences learned in context. *Journal of Marketing Research* **45**, 145–158.

Ashraf N., D. Karlan and W. Yin 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. *Quarterly Journal of Economics* **121**, 635–672.

Atran S., H. Sheikh and A. Gomez 2014. Devoted actors sacrifice for close comrades and sacred cause. *Proceedings of the National Academy of Sciences USA* **111**, 17702–17703.

Austad S.N. 1997. Postreproductive survival. In *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academies Press eBooks.

Barker A. 1977. Why did Socrates refuse to escape? *Phronesis* **22**, 13–28.

Baumeister R.F. and S.J. Scher 1988. Self-defeating behavior patterns among normal individuals: review and analysis of common self-destructive tendencies. *Psychological Bulletin* **104**, 3–22.

Becker G.S. 1976. *The Economic Approach to Human Behavior*. Amsterdam: Amsterdam University Press.

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

Bergh A. and P.C. Wichardt 2018. Accounting for context: separating monetary and (uncertain) social incentives. *Journal of Behavioral and Experimental Economics* **72**, 61–66.

Biesalski H.K. 2005. Meat as a component of a healthy diet – are there any risks or benefits if meat is avoided in the diet? *Meat Science* **70**, 509–524.

Bonein A. and L. Denant-Boèmont 2015. Self-control, commitment and peer pressure: a laboratory experiment. *Experimental Economics* **18**, 543–568.

Bostrom N. 2012. The superintelligent will. *Minds and Machines* **22**, 71–85.

Brandtstädter J. and R.M. Lerner (eds) 1999. *Action and Self-Development: Theory and Research Through the LifeSpan*. London: Sage.

Burke P.J. 2006. Identity change. *Social Psychology Quarterly* **69**, 81–96.

Button M.E. 2018. Bounded rationality without bounded democracy: nudges, democratic citizenship, and pathways for building civic capacity. *Perspectives on Politics* **16**, 1034–1052.

Bühler C. 1964. The human course of life in its goal aspects. *Journal of Humanistic Psychology* **4**, 1–18.

Camerer C. 2003. *Behavioral Game Theory*. Princeton: Princeton University Press.

Carlsson F., O. Johansson-Stenman and P.K. Nam 2014. Social preferences are stable over long periods of time. *Journal of Public Economics* **117**, 104–114.

Chulef A.S., S.J. Read and D.A. Walsh 2001. A hierarchical taxonomy of human goals. *Motivation and Emotion* **25**, 191–232.

Cohen J.D., S.M. McClure and A.J. Yu 2007. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences* **362**, 933–942.

Colin-Jaeger N. and M. Dold 2025. Individual autonomy and public deliberation in behavioral public policy. *Humanities and Social Sciences Communications* **12**, Article no. 430 (2025).

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

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

Conway M. and H. Williams 2008. Autobiographical memory. In *Learning and Memory: A Comprehensive Reference*, eds J.H. Byrne et al., 893–909. Oxford: Elsevier.

Cote J., A.N. Dreiss and J. Clobert 2008. Social personality trait and fitness. *Proceedings of the Royal Society B: Biological Sciences* **275**, 2851–2858.

Cowan N. 2010. The magical mystery four. *Current Directions in Psychological Science* **19**, 51–57.

Deci E.L. and R.M. Ryan 2000. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychological Inquiry* **11**, 227–268.

DellaVigna S. 2009. Psychology and economics: evidence from the field. *Journal of Economic Literature* **47**, 315–372.

Delmotte C. and M. Dold 2022. Dynamic preferences and the behavioral case against sin taxes. *Constitutional Political Economy* **33**, 80–99.

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

Dold M., E. Van Emmerick and M. Fabian 2024. Taking psychology seriously: a self-determination theory perspective on Robert Sugden's opportunity criterion. *Journal of Economic Methodology* 1–18. https://doi.org/10.13140/RG.2.2.12287.43685/1.

Durizzo K., K. Harttgen, F. Tediosi *et al.* 2022. Toward mandatory health insurance in low-income countries? An analysis of claims data in Tanzania. *Health Economics* **31**, 2187–2207.

Erlangga D., M. Suhrcke, S. Ali and K. Bloor 2019. The impact of public health insurance on health care utilisation, financial protection and health status in low- and middle-income countries: a systematic review. *PLoS ONE* **14**, e0219731.

Fabian M. and M. Dold 2025. Agentic preferences: a foundation for nudging when preferences are endogenous. *Behavioural Public Policy* **9** 1–21.

Ford M.E. and D.H. Ford 2019. *Humans as Self-Constructing Living Systems: Putting the Framework to Work*. London: Routledge.

Giné X., D. Karlan and J. Zinman 2010. Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics* **2**, 213–235.

Grahek I., X. Leng, S. Musslick and A. Shenhav 2023. Control adjustment costs limit goal flexibility: empirical evidence and a computational account. *bioRxiv*.

Gruber J.H. and S. Mullainathan 2005. Do cigarette taxes make smokers happier. *The B.E. Journal of Economic Analysis & Policy* **5**(1). https://doi.org/10.1515/1538-0637.1412.

Gosling S.D. 2008. Personality in non-human animals. *Social and Personality Psychology Compass* **2**, 985–1001.

Grimm S.R. 2008. Epistemic goals and epistemic values. *Philosophy and Phenomenological Research* **77**, 725–744.

Gugerty M.K. 2007. You can't save alone: commitment in rotating savings and credit associations in Kenya. *Economic Development and Cultural Change* **55**, 251–282.

Habermas T. and S. Bluck 2000. Getting a life: the emergence of the life story in adolescence. *Psychological Bulletin* **126**, 748–769.

Hamilton W.D. 1964. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology* **7**, 17–52.

Hausman D.M. and B. Welch 2009. Debate: to nudge or not to nudge. *Journal of Political Philosophy* **18**, 123–136.

Häußermann J.J. 2019. Nudging and participation: a contractualist approach to behavioural policy. *Philosophy of Management* **19**, 45–68.

Hicks J.A. and L.A. King 2009. Positive mood and social relatedness as information about meaning in life. *Journal of Positive Psychology* **4**, 471–482.

Hooker S.A., K.S. Masters and C.L. Park 2018. A meaningful life is a healthy life: a conceptual model linking meaning and meaning salience to health. *Review of General Psychology* **22**, 11–24.

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

Jacobs D.R. and L.C. Tapsell 2013. Food synergy: the key to a healthy diet. *Proceedings of the Nutrition Society* **72**, 200–206.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

Kaplan H. 1997. The evolution of the human life course. In *Between Zeus and the Salmon: The Biodemography of Longevity*. National Academies Press eBooks.

Kaplan H. and S.W. Gangestad 2015. Life history theory and evolutionary psychology. In *The Handbook of Evolutionary Psychology*, 68–95. Chichester: Wiley.

Kast F., S. Meier and D. Pomeranz 2018. Saving more in groups: field experimental evidence from Chile. *Journal of Development Economics* **133**, 275–294.

Kemper F. and P.C. Wichardt 2024a. Welfare justifications and responsibility in political decision making – the case of nudging. *Critical Policy Studies*, 1–17.

Kemper F. and P.C. Wichardt 2024b. Procedurally justifiable strategies: integrating context effects into multistage decision making. *Review of Behavioral Economics* **11**, 313–347.

King L.A. and J.A. Hicks 2007. Whatever happened to "what might have been"? regrets, happiness, and maturity. *American Psychologist* **62**, 625–636.

Kopczuk W. and J. Slemrod 2003. Dying to save taxes: evidence from estate-tax returns on the death elasticity. *The Review of Economics and Statistics* **85**, 256–265.

Leggett W. 2014. The politics of behaviour change: nudge, neoliberalism and the state. *Policy & Politics* **42**, 3–19.

Loewenstein G., L. Thompson and M.H. Bazerman 1989. Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology* **57**, 426–441.

McAdams D.P. and J.L. Pals 2006. A new Big Five: fundamental principles for an integrative science of personality. *American Psychologist* **61**, 204–217.

McAvoy A., J. Kates-Harbeck, K. Chatterjee and C. Hilbe 2022. Evolutionary instability of selfish learning in repeated games. *PNAS Nexus* **1**(4).

McEwen B.S. 2008. Central effects of stress hormones in health and disease: understanding the protective and damaging effects of stress and stress mediators. *European Journal of Pharmacology* **583**, 174–185.

Meyer C., L. Taranis, H. Goodwin and E. Haycraft 2011. Compulsive exercise and eating disorders. *European Eating Disorders Review* **19**, 174–189.

Mroczek D.K. and A. Spiro 2003. Modeling intraindividual change in personality traits: findings from the normative aging study. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **58**, P153–P165.

Nettle D. 2006. The evolution of personality variation in humans and other animals. *American Psychologist* **61**, 622–631.

O'Grady T. 2017. How do economic circumstances determine preferences? evidence from long-run panel data. *British Journal of Political Science* **49**, 1381–1406.

Ojanen T., K. Aunola and C. Salmivalli 2007. Situation-specificity of children's social goals: changing goals according to changing situations? *International Journal of Behavioral Development* **31**, 232–241.

Polak R., E.M. Phillips and A. Campbell 2015. Legumes: health benefits and culinary approaches to increase intake. *Clinical Diabetes* **33**, 198–205.

Raggatt P. 2006. Putting the Five-Factor model into context: evidence linking big five traits to narrative identity. *Journal of Personality* **74**, 1321–1348.

Rains S.A. 2012. The nature of psychological reactance revisited: a meta-analytic review. *Human Communication Research* **39**, 47–73.

Restrepo O.N. and S. Vaisey 2024. Opinions on hard-to-discuss topics change more via cohort replacement. *Evolutionary Human Sciences* **6**, e25, 1–18.

Roberts B.W. and R.W. Robins 2000. Broad dispositions, broad aspirations: the intersection of personality traits and major life goals. *Personality and Social Psychology Bulletin* **26**, 1284–1296.

Roberts B.W. and D. Mroczek 2008. Personality trait change in adulthood. *Current Directions in Psychological Science* **17**, 31–35.

Rokeach M. 1973. *The Nature of Human Values*. Free Press.

Ruegsegger G.N. and F.W. Booth 2017. Health benefits of exercise. *Cold Spring Harbor Perspectives in Medicine* **8**(7), a029694.

Samuelson P.A. 1938. A note on the pure theory of consumer's behaviour. *Economica* **5**, 61–71.

Samuelson P.A. 1948. Consumption theory in terms of revealed preference. *Economica* **15**, 243–253.

Schütze T., C. Spitzer and P. Wichardt 2025. Nudging: an experiment on transparency, controlling for reactance and decision time. *Journal of Economic Psychology* **107**, 102797.

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press

Schwartz S.H. 1992. Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology* **25**, 1–65.
Schwartz J., D. Mochon, L. Wyper *et al.* 2014. Healthier by precommitment. *Psychological Science* **25**, 538–546.
Sen A. 1985. Goals, commitment, and identity. *Journal of Law, Economics, and Organization* **1**, 341–355.
Shim S., L. Dorworth, J. Lasrado and C. Santerre 2004. Mercury and fatty acids in canned tuna, salmon, and mackerel. *Journal of Food Science* **69**(9). https://doi.org/10.1111/j.1365-2621.2004.tb09915.x.
Sillaber I. and M.S.H. Henniger 2004. Stress and alcohol. *Annals of Medicine* **36**, 596–605.
Simler K. and R. Hanson 2017. *The Elephant in the Brain: Hidden Motives in Everyday Life*. Oxford: Oxford University Press.
Špecián P. 2019. The precarious case of the true preferences. *Society* **56**, 267–272.
Špecián P. 2022. *Behavioral Political Economy and Democratic Theory: Fortifying Democracy for the Digital Age*. London: Routledge.
Sugden R. 2017. Do people really want to be nudged towards healthy lifestyles? *International Review of Economics* **64**, 113–123.
Suk J. 2011. Preventive health at work: a comparative approach. *American Journal of Comparative Law* **59**, 1089–1134.
Sunstein C.R. 2018. "Better off, as judged by themselves": a comment on evaluating nudges. *International Review of Economics* **65**, 1–8.
Tabibnia G., A.B. Satpute and M.D. Lieberman 2008. The sunny side of fairness. *Psychological Science* **19**, 339–347.
Talevich J.R., S.J. Read, D.A. Walsh, R. Iyer and G. Chopra 2017. Toward a comprehensive taxonomy of human motives. *PLoS ONE* **12**(2), e0172279.
Tirole J. 2002. Rational irrationality: some economics of self-management. *European Economic Review* **46**, 633–655.
Thaler R.H. 2015. *Misbehaving: The Making of Behavioral Economics*. W.W. Norton & Company.
Thaler R.H. and H.M. Shefrin 1981. An economic theory of self-control. *Journal of Political Economy* **89**, 392–406.
Thaler R.H. and S. Benartzi 2004. Save More Tomorrow™: using behavioral economics to increase employee saving. *Journal of Political Economy* **112**, 164–187.
Thaler R.H. and C.R. Sunstein 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.
Tversky A. and D. Kahneman 1986. Rational choice and the framing of decisions. *Journal of Business* **59**, 251–278.
Twigg J.L. 1999. Obligatory medical insurance in Russia: the participants' perspective. *Social Science & Medicine* **49**, 371–382.
Whitman D.G. and M.J. Rizzo 2015. The problematic welfare standards of behavioral paternalism. *Review of Philosophy and Psychology* **6**, 409–425.
Zakharenko R. 2016. Nothing else matters: evolution of preference for social prestige. *Mathematical Social Sciences* **80**, 58–64.

Petr Krautwurm is a PhD candidate at the Prague University of Economics and Business. His interdisciplinary research focuses on the area of philosophical and methodological questions about economics; such as second-order preferences, precommitments, and implications for paternalism.

Philipp C. Wichardt is a Professor of Economics at the University of Rostock and a research fellow of the CESifo Munich and the Kiel Institute for the World Economy. His interdisciplinary research focuses on the social and psychological determinants of individual decision-making. It has been published in various different fields of the social sciences.

Cite this article: Krautwurm P and Wichardt PC (2026). Preferences, goals and implications for paternalism. *Economics and Philosophy* **42**, 154–173. https://doi.org/10.1017/S0266267125100436

https://doi.org/10.1017/S0266267125100436 Published online by Cambridge University Press