Nudges for People who Think

Aba Szollosi[1]*, Nathan Wang-Ly[2], & Ben R. Newell[2,3]*

[1] Department of Psychology, University of Edinburgh

[2] School of Psychology, UNSW Sydney

[3] UNSW Institute for Climate Risk & Response

Author Note

The support of the Australian Research Council (DP190101076; DP190101675) is gratefully acknowledged.

*These authors contributed equally. Address correspondence to:

Aba Szollosi, Department of Psychology, University of Ediburgh, EH8 9JZ, UK, Email: aba.szollosi@gmail.com

Ben R. Newell, School of Psychology, UNSW Sydney, NSW 2052, Australia. Email: ben.newell@unsw.edu.au

Abstract

The naiveté of the dominant 'cognitive-miser' metaphor of human thinking hampers theoretical progress in understanding how and why subtle behavioral interventions – 'nudges' – could work. We propose a reconceptualization that places the balance in agency between, and the alignment of representations held by, people and choice architects as central to determining the prospect of observing behaviour change. We argue that two aspects of representational (mis)alignment are relevant: cognitive (how people construe the factual structure of a decision environment) and motivational (the importance of a choice to an individual). Nudging thinkers via the alignment of representations provides a framework that offers theoretical and practical advances and avoids disparaging people's cognitive capacities.

Keywords: behavior change, choice architecture, cognitive miser, decision making, nudge, representation alignment

It is now commonplace to hear that addressing the grand challenges of today's society requires substantial changes in behaviour (e.g., Bergquist, Thiel, Goldberg, & van der Linden, 2023; Van Bavel et al., 2020). This call to arms places psychological and behavioural science at the forefront of understanding how such widespread change can be achieved. Answering that call has led many researchers to make bold claims for the potential of simple techniques that facilitate positive behaviour change without impinging on people's freedom of choice (Thaler & Sunstein, 2008). These techniques, collectively and colloquially known as 'nudges' capitalize on promoting 'desirable' options by making changes to the choice architectures (physical, social, and psychological) in which decisions are made.

Discussions about the effectiveness of nudges are receiving increasingly widespread coverage in the academic literature as well as in wider general discourse (Chater & Loewenstein, 2022; Hallsworth, 2022; Simmons, Nelson, & Simonsohn, 2022). Such debate is welcome and necessary given the importance of the challenges we must address and the potential for the low-cost, quick-win policy instruments that nudges seemingly represent. However, much of this recent debate has centered on questions about 'what works and when does it work' at the expense of questions about *how* and *why* (e.g., Osman et al., 2020; Szaszi et al., 2018). Such focus is understandable; a government that wants rapidly to encourage people to wear masks or stay at home may not care all that much about why a particular nudge works, it just wants to know that it will work, and quickly.

In the long run, however, this approach is self-defeating: if we do not understand why a technique works and then it stops working, we are unlikely to know how to make it work again. Without a deeper understanding, we cannot know how long a particular nudge will work – or if it would work at all in a novel situation (i.e., that is not the unique one it was originally observed in).

Offering these techniques to governments without such an understanding may even undermine the credibility of behavioural scientists. Despite a general sense that there is something called "Nudge Theory" (e.g., https://en.wikipedia.org/wiki/Nudge_theory), the promise of such a theory seems like a mirage given the assortment of often loosely defined and tangentially related intervention techniques in the choice architect's toolbox. This state of affairs can lead to a "throw everything at the wall and see what sticks" approach to testing nudges, which substantially constrains how general our understanding of them can become[1].

To overcome this limited approach, here we offer a simple unified theoretical perspective of how and why nudges could work. Central to this perspective is challenging the cognitive miser metaphor that pervades the theorizing about nudging – that people are lazy, capacity-limited, and largely unaware of the reasons underlying their behavior. This metaphor invites a passive-actor framing, but it is over-simplified and unjustified. In order to advance our understanding of how and why nudges work, we need to take more seriously the idea that people are *thinkers* – active agents in determining their behaviour and not passive recipients unthinkingly swayed by subtle changes in the environment.

We build our case by 1) reviewing the deficiencies in the cognitive miser metaphor, 2) arguing that methods currently promoted in the field (such as meta-analyses) cannot overcome these shortcomings, and 3) presenting a novel theoretical framework that can overcome the limitations of the prevailing approach, but that also reveals strong boundaries on when we can

[1] See: https://www.jasoncollins.blog/posts/megastudy-scepticism for an interesting discussion of this issue.

expect nudges to work. We conclude with a discussion of the practical consequences of this framework for future research on behavioural interventions.

Deficiencies in the cognitive miser metaphor

The theoretical foundations of nudges are derived from two stylized ideas about human thinking. The first is that all else equal, we tend to avoid mental effort. Thinking is costly and thus aversive and so if we can conserve our mental resources we will (Kahneman, 2011). The second is the idea that mental processes can be dichotomized and compartmentalised into different boxes that house styles of thinking that in some way align or coalesce (Evans & Stanovich, 2013). There are as many versions of this idea as there are papers written about it, but broadly speaking one box, 'System 1', captures automatic thinking that tends to be fast, associative, and operate independently of working-memory. The other box, 'System 2', is the slower, deliberative system that relies on rules, is capacity-limited and cognitively effortful.

The combination of these two ideas – that effortful thinking is aversive, but that we have a system that can think automatically – leads to the seductive notion that people's behaviour can be influenced via interventions that operate on System 1, while bypassing System 2 altogether. If done successfully such interventions offer the promise of achieving what is often perceived as effortful behaviour change via an effortless channel. For example, rather than exhorting people to eat low-fat foods via information campaigns, just place those foods in more convenient places in supermarkets or canteens, where they are more likely to be 'unthinkingly' chosen.

Despite claims that dual-system dichotomies should only be considered metaphors, or devices to help organise thinking about thinking (Chater, 2018), the general framework and its combination with the notion of humans as cognitive misers has become widespread. The idea that we need to influence the unthinking ‘lazy’ part of our cognitive apparatus in order to achieve change has garnered proponents from the World Bank (2015) to the National Academy of Sciences (2015) as well as countless consultancies promoting their tools for capitalizing on the powers of automatic thinking.

There are several reasons to be sceptical about this characterization of human thinking. The claim that mental effort is universally aversive is readily countered by evidence of engagement in cognitive activities that require apparently costly thinking. From solving the daily Wordle, to completing crosswords, Sudokus, and playing computer and strategy games like chess, it is clear that people like to think and challenge their mental capacity under the right circumstances (Embrey, Donkin, & Newell, 2023; Inzlicht, Shenhav, & Olivola, 2018; Thomson & Oppenheimer, 2022). The discrepancy between avoiding and seeking mental effort presumably arises because the goals or motivation for pursuing effortful activities can differ both within individuals (e.g., across time or tasks) and between individuals. This heterogeneity suggests that an approach like nudging which is predicated on the blanket assumption that people will choose to avoid effort is likely to be insufficient. Just as there is heterogeneity in preference for effort so there will be heterogeneity in the fit between people and techniques designed to prompt engagement or circumvent thinking (Bryan, Tipton, & Yeager, 2021).

The evidence for the automaticity of thinking is on similarly shaky foundations. Canonical demonstrations of tasks that apparently do not require the involvement of working memory, or deliberative thought have been challenged rigorously (Newell & Shanks, 2014; 2023). To take one prominent example, the idea that complex decisions (e.g., which apartment or car to buy) can be solved by allowing unconscious thought processes to supersede capacity-limited conscious thought has been resolutely refuted (Nieuwenstein et al., 2015).

Even more striking is the volte-face in opinion about the kinds of behaviours that can be influenced in the absence of awareness. Much of subtle behaviour change-intervention is predicated on the questionable notion that we are often unaware of the reasons why we behave as we do. This view, often attributed to the seminal work of Nisbett and Wilson (1977), has led to the proliferation of claims about the power that subtle environmental cues can have on a range of behaviours (Bargh, 2017). While many of the studies on which these claims are based remain prominent in practitioners' handbooks and recommendations (e.g., Dolan et al., 2010), the academic literature is now much more sanguine about the reliability and relevance of this literature (Newell & Shanks, 2023; Ritchie, 2020). In large part this shift in opinion has been driven by failures to replicate landmark studies that purported to find evidence for unconscious influences on our behaviour (e.g., O'Donnell et al., 2018; Shanks et al., 2013, 2015). Faced with this mounting evidence, the idea that higher-level cognitive activities of the kind usually targeted by nudges – judgments, decisions, choices – are strongly influenced by factors outside of our awareness has become less and less tenable.

The major concern then is that the current theoretical underpinnings of nudges are based on literatures that are at worst outdated and irrelevant and at best admitting of much more nuanced interpretations. The idea that people are cognitively lazy and act 'unthinkingly' does not bear scrutiny. Our contention is that this misguided application of the cognitive-miser metaphor in attempting to understand 'why nudges work' has dominated – and paralysed – the current behavioural intervention-approach. The main reason for this paralysis is that the metaphor is vague and flexible and thus offers little in the way of clear predictions about when or why a technique might work.

In order to overcome the ‘do they/don’t they work’ question, we need to abandon the current metaphor and re-think how nudges *could* work in a world where people are active agents in determining their behaviour and not passive unthinking recipients of information who are buffeted around within a pre-determined choice architecture. However, before pursuing this line of argument we first consider another approach for assessing the effectiveness question – meta-analyses.

The what and when but not the ‘why’

Meta-analyses are useful tools for identifying the kinds of nudges that are effective, and the situations in which they are effective, but they shed little light on the reasons why techniques fail or succeed. For example, a recent meta-analysis by Jachimowicz and colleagues (2019) compared default nudges – where one option is pre-selected for people – across consumer choice, environmental, and health domains. They found that defaults have larger effects in consumer choice contexts than environmental contexts, and stated that *“We can only speculate about why this occurs ... Perhaps consumer preferences are less strongly held than preferences in other domains and environmental preferences more strongly”* (p.176). This conclusion suggests that the domain of the default nudge is not really the important factor in its efficacy, but rather some other dimension (in this case, strength of preference) that offers greater explanatory power.

A similar conclusion can be drawn from a recent large scale meta-analysis of choice architecture interventions. Mertens and colleagues (2022) analyzed data from over 200 studies across a range of intervention techniques and domains concluding that overall, interventions promote behaviour change with a small to medium effect size (*Cohen’s d* = 0.45). They categorized techniques as targeting *decision information, decision assistance,* and *decision structure.* A key finding of the meta-analysis was that *structure* interventions, those which aim to make it simpler to choose ‘better’ options via the setting of defaults or removal of ‘frictions’, were more consistently effective at promoting behaviour-change than the *information* or *assistance* techniques. Mertens and colleagues speculate that the higher effectiveness of the decision-structure interventions may be due both to the lower demand on information processing of the structure relative to the information and assistance techniques and the fact that they are less likely to engage people in deliberative assessment of the goals and values relevant for a particular decision. To illustrate this last point, the authors suggest that nutritional labels – a popular decision information technique – are more effective for consumers who are already concerned about their health than those who are not. In essence, Mertens et al., argue that *information* and *assistance* techniques are likely to be more susceptible to the heterogeneity in values and goals that might be triggered when people are confronted with such interventions, thereby potentially weakening their overall effectiveness in the general population.

This conclusion is important, albeit still constrained by the desire to compartmentalize interventions into automatic or deliberative boxes. Its importance lies in highlighting an under-appreciated aspect of *why* different techniques might work: namely the degree to which any attempt to change behaviour can be construed as a social interaction between the person who is attempting to invoke that change (the choice architect) and the person whose behaviour is being targeted (de Ridder et al., 2022; McKenzie et al., 2018; Krijnen et al., 2017; Sher et al., 2022).

But a meta-analytic approach can get us only so far in the pursuit of understanding why this interaction is important. This is partly because meta-analyses are limited by the quality of their inputs; failure to include studies that do not see the light of day because they did not ‘work’ (a publication bias) can lead to overestimates of the effectiveness of nudges. Indeed, critics argue that when statistically controlling for this problem, the purported effects found in the meta-analysis fully disappear or are at least substantially reduced (Maier et al., 2022; Szaszi et al., 2022). Such statistical debates notwithstanding, meta-analyses cannot reveal the reasons why different techniques might flourish or fail, because they only offer us a snapshot of what worked at some point in the (recent) past – thus limiting their ability to contribute to the "nudge theory" problem. To advance our understanding, we need to build theories that incorporate values, goals, and preferences and abandon the effort-avoiding, automaton passive-actor framing.

Cognitive misers vs active thinkers: Tipping the balance of agency

Nudge research has relied on the cognitive miser metaphor for a theoretical foundation, according to which people's purported tendency for laziness leads to systematic cognitive biases – a set of behaviours deviating from what researchers expect as optimal. Nudges supposedly exploit these biases, which serve as effortless System 1 pathways effecting effortful behaviour change. Although relying on this framework may be intuitively appealing as people's phenomenological experiences often map onto the vague labels of the cognitive miser metaphor, the many shortcomings that we have reviewed previously naturally raise the question of whether a better alternative exists.

Here we outline and argue in favour of such an alternative, which frames people as active thinkers whose aim is to understand their environments increasingly better in order to make better decisions (Szollosi & Newell, 2020). Under this view, people deviate from researchers' expectations not because they are hardwired to be biased, but because their respective representations of the choice environment are misaligned – they do not construe the choice that they are facing similarly. A crucial difference from the cognitive miser view is that these misalignments (mistakes or biases) are not systematic[2]: representations can become better aligned through processes that we argue are responsible for the effects of existing nudges and that can be built on when designing future ones.

For our current purposes two ways of representational (mis)alignment are relevant: cognitive and motivational. This roughly means that people can either construe the factual structure of a decision environment differently (cognitive) or they can think differently about the importance of a particular choice (motivational). More concretely, when people make mistakes on tasks often held up as examples of the biased nature of people's cognition (Tversky & Kahneman, 1974) the mistakes are occurring either a) because participants construe the task differently from the experimenter – for instance, due to not having the requisite understanding of probabilistic principles (Koehler, 1996) or making use of other task-relevant information the experimenter failed to consider (Sher & McKenzie, 2006); or b) because they do not find the task particularly interesting or worthy to engage with at a more substantial level than responding randomly or superficially. Note that this is not a general tendency to be lazy, but an active inferential process about what they find interesting, and whether solving the task properly fits within their goals – some people may find it interesting to solve these tasks well.

These two ways in which representations can be misaligned are important for nudge research, as together they provide a more unified way to understand how interventions can work and so a more robust framework for building them. Notably, this framing espouses many of the

[2] As people build their representations in a substantial part from background knowledge that is stable over longer timeframes (e.g., common-sense cultural knowledge, material taught in schools, etc.), the mistakes that they make can be similar across people and over extended time periods, which can contribute to the appearance of pervasiveness. However, just because a representation only changes over longer timeframes does not mean that it should be considered as a stable or systematic feature of people's decision-making processes (Szollosi, Donkin, & Newell, 2023).

problems stemming from the cognitive miser metaphor: people are not hardwired to be cognitively lazy (although they can be when they are unmotivated) and biased (although they can make correctable mistakes). As such, their behaviour can be changed not by exploiting these inborn biases, but through methods that treat them as actors with high agency, dovetailing with recent suggestions on the topic (de Ridder et al., 2022; Hertwig & Grüne-Yanoff, 2017; McKenzie et al., 2018; Krijnen et al., 2017; Sher et al., 2022).

Nudging thinkers by aligning representations

The active thinker framing of the researcher–participant relationship readily applies to the choice architect–decision maker relationship as well. Specifically, when the default behaviour of a decision maker is considered suboptimal by the choice architect, it can be the result of the decision maker's different factual understandings of the decision environment, their motivation to achieve different goals in that environment, or a mixture of both. What needs to happen for behaviour change to occur is for the relevant aspects of the representation to become better aligned between the architect and the decision maker – a process that can occur via a range of different pathways.

Although representational alignment develops primarily through relatively slow and effortful means such as education and the broader cultural transmission of knowledge, there are simpler, low-cost pathways through which nudges can increase it albeit by a much more limited extent. Such interventions can influence behaviour through three main pathways (illustrated on Figure 1): a) they can work by serving as *reminders* of behavioural goals that the decision maker already had but momentarily forgot about; b) they can work through the *persuasion* of the decision maker that the recommended behaviour is beneficial for them; and c) they can work by *implying* *threat* to the decision maker if they do not follow the recommendations of the choice architect. These pathways characterise ways in which simple methods can result in behaviour change while retaining the decision maker's agency, and they are also consistent with popular descriptive models of behaviour change (e.g., Duckworth, Milkman, & Laibson, 2018; Michie, Van Stralen, & West, 2011).

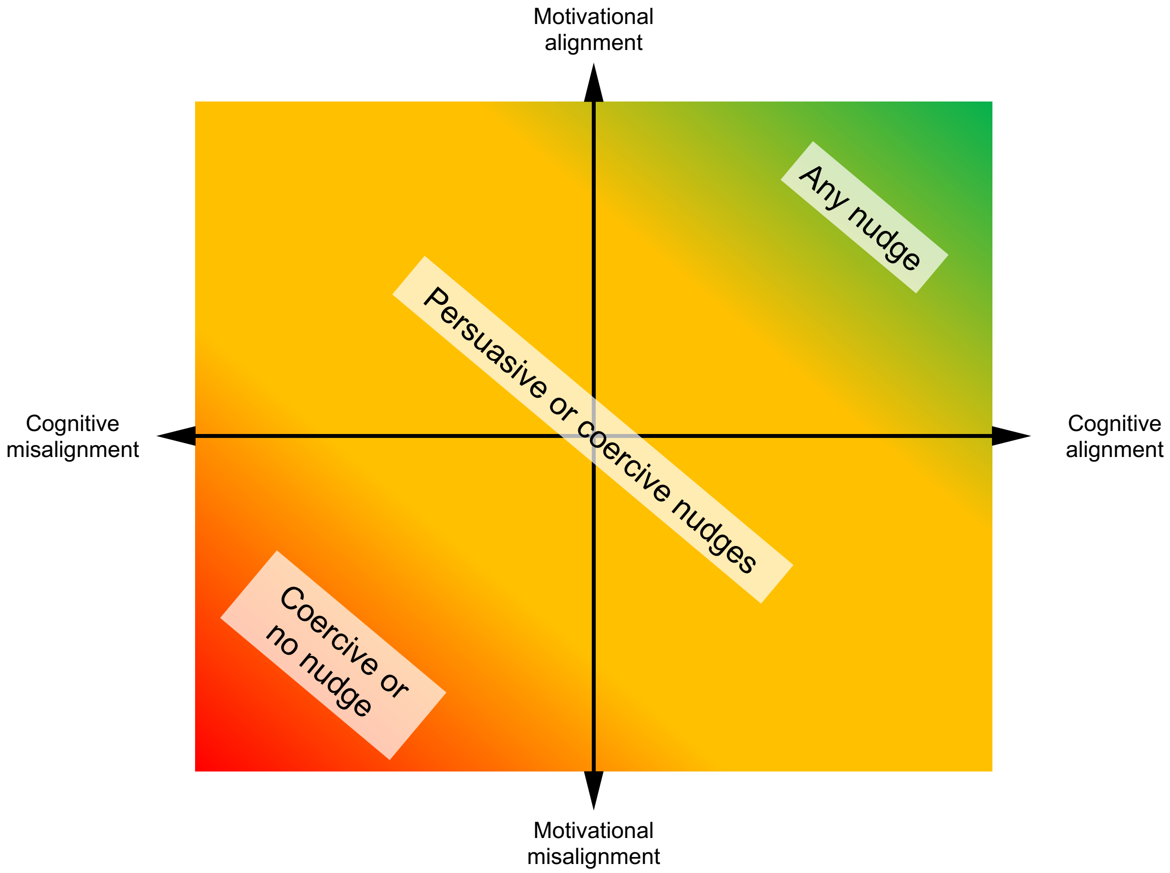

Figure 1. Illustrative classification of the types of nudges that can be expected to work under varying degrees of representational alignment.

A choice architect can use any nudge to influence the decision maker's behaviour when their representations are highly aligned, both at the cognitive and motivational levels (green region on Figure 1). As the most likely source of unaligned behaviour under such conditions is that inattention temporarily misaligned representations, any nudge can remind the decision maker of their previously set goals. The domain of healthy food choice provides good examples (e.g., Bucher et al., 2016). In this case, we can expect nudges to work when the decision maker agrees with the choice architect on what a healthy choice is and wants to generally eat healthy, but temporarily forgets about this (perhaps not strongly formed) goal. From this perspective we can also derive the expectation that such a nudge will not work when representations are misaligned either cognitively or motivationally. For example, if the decision maker does not share the representation of what constitutes healthy or unhealthy food, or what effects the consumption of those foods might have on their health, we should not expect this type of nudge to influence their behaviour. Similarly, independent from whether they share the representation at the cognitive level, we should also not expect nudges to work when the decision maker does not think that healthy eating behaviour is important or is a desirable goal.

In these latter cases, when representational alignment is somewhat lower (orange region on Figure 1), a way in which nudges can still work is by persuading the decision maker. Examples of this include employer recommendations for pension schemes or health insurance providers (Benartzi et al., 2017). These nudges can work through increasing either cognitive or motivational alignment. For example, cognitively, they can make the decision maker recognise and address a mistaken or neglected aspect of their representation (e.g., by trusting that the choice architect making the recommendation had their best interest in mind when they did not have time to evaluate every option – or, when there is no trust or when the stakes are high, by prompting a more detailed evaluation of the options). Motivationally, such nudges can make people recognise the importance of a problem that they previously may not have considered a problem (e.g., by making younger people aware of the existence and importance of pension schemes).

When the level of representational alignment is low (red region on Figure 1) – because the decision maker strongly disagrees on facts, importance, or both – only nudges that imply threat can work. For example, sending personalised letters about unpaid taxes (e.g., Hallsworth et al., 2017) can imply an immediate threat of legal action (compared to when those letters are more generically phrased). In other cases, when the nudge cannot be interpreted as a threat and representational content is not shared, the nudge will fail to work. For example, most food choice nudges cannot evoke a threatening interpretation and as such will not overcome such strong misalignment of representations when they exist.

The limits of nudges

Understanding how nudges work through representational alignment also illuminates the limits of what they can be expected to achieve. Specifically, it identifies the limits of nudges as people's actual understanding of and/or goals within their environments. Although these might sound self-evident or trivial concerns, as representational content and motivation is determined by a multitude of factors (e.g., cultural knowledge, people's own knowledge and experiences, situational elements, see Footnote 2.) the space of shared representational content between choice architects and decision makers quickly diminishes with increasing representation variability. This is particularly relevant for representations about contentious or personally important issues that nudges are often aimed at (health, financial, and environmental behaviours), because they can and do vary greatly across people both at cognitive and motivational levels.

Another limit this view highlights is for meta-analytic (Mertens et al., 2022) and large-scale experimental (Milkman et al., 2021) methods in nudge research, as foreshadowed earlier. This is because these methods can only show a snapshot of how effectively a nudge *currently* works, but when the representations that underlie the behaviour change, the nudge might stop working generally. Note that this criticism holds irrespective of how well-conducted the meta-analysis was (cf., Meier et al., 2022; Szaszi et al., 2022). For example, imagine that using the currently available best meta-analytic technique we estimate that a specific nudge has a reliably medium effect-size on healthy food choice. In what sense should we expect this nudge to work generally? People a couple hundred years ago would have found such nudges meaningless, as their representation of their choice environment primarily centred around finding and consuming any available food. People a couple hundred years from now might represent their environment differently for other reasons: perhaps new research will have identified other types of food as healthy and unhealthy, or perhaps medicine will have advanced to a degree where diseases resulting from unhealthy food can be cured with ease. These somewhat outlandish examples illustrate the transient nature of people's representations well, and – more relevant for practitioners aiming to change current-day behaviour – it is not hard to think of representations that have changed on much shorter time frames (e.g., the rapidly changing views on the benefits of mask use, lockdowns, or vaccinations during the COVID-19 pandemic). In none of these cases should we expect the nudge to have the same effect it once produced in a mega-study or meta-analysis under the different respective representations. Instead, under the view we outlined, the results of mega-studies and meta-analyses can only be taken into consideration after the possibility of representational misalignment is accounted for (e.g., by monitoring representational change over time, or documenting current differences between cultures).

Practical consequences for future research

Our theoretical reassessment of how and why nudges can be expected to work has far-reaching practical relevance for future nudge research and potentially even for broader behaviour change research. Through an example about pension decisions, we illustrate how considering the extent of representational alignment can help derive expectations about when and what kind of nudges could work. There are two main ways in which we can deploy this theory in practice: 1) assessing representational alignment and 2) developing nudges based on that assessment.

Assessing the extent of the target population's representational alignment can serve as a useful starting point. This could proceed through asking a sample of people we aim to nudge about both the cognitive and motivational aspects of their representation (e.g., via a survey and/or relatively unstructured interviews, but in some cases, we can infer important dimensions on a theoretical basis), with the aim of identifying whether they share the underlying representation of the decision problem and how important they consider the choice.

For pension decisions, we may identify as the most relevant cognitive dimensions for people's decisions whether people understand how pension funds work and how feasible it would be for them to make contributions. That is, decision makers might have a clear understanding of the various funds and schemes they could contribute to or they might not – and similarly they may or may not think that they have sufficient income to make contributions. Note that the idea that they have a decent understanding of options and the means to contribute (i.e., that their representation of the factual side of the problem is the same as the choice architect's) is often implicit in how nudges are currently developed.

On the motivational side, personal importance of retirement saving and the reasons underpinning that goal might be important dimensions. For example, some of the people who do not share this goal with the choice architect may not want to save because they have not thought about it in detail, while others may think it is too far in the future and they can start later, or have a different conception about how much is ‘enough’ for retirement (Goldin, Homonoff, Patterson, & Skimmyhorn, 2020). Note, again, that the assumption that the decision maker’s goal is in alignment with the architect’s goal is often implicit in current nudge development.

With the extent of representational alignment assessed, we can now start selecting particular nudges to use. From our earlier discussion we know that the more the representations align the more we can expect nudges to work. For people with highly aligned representations – those who understand how pensions work, have the means, and want to contribute – any type of nudge would work, since if they are not already contributing it is presumably due to inattention. For people misaligned in their factual understanding of – for example, if they do not have the relevant information – persuasive nudges might work. For example, employer ranking of various pension fund options could help people gather more information for their choice. People misaligned on other dimensions – people who do not have the means or specifically do not want to contribute – likely cannot be influenced through nudges unless they have coercive implications (such as government mandated enrolment).

There are several possible extensions of this method. One relatively easy idea is to transfer the knowledge gained from developing nudges for a target population this way to another population by using clustering-based methods (e.g., Deetlefs et al., 2019). For instance, we can identify and measure the frequency of various representational clusters in our initial population and how well our various nudges worked for each of them. This can help estimate how effective the nudge might be in another population after only measuring the representational base rates. Another extension could be to individually develop nudges for people who are not aligned. Although the more generic nudges considered above might work for most of them, people often have immensely variable reasons for diverging from the understanding of the choice architect. While this extension can make the method more costly, increasing the resolution of understanding regarding these individual reasons could also increase the efficacy of the nudges (cf., Bryan et al., 2021; Osman et al., 2020). Lastly, misalignment can also be the result of the choice architect's inaccurate representation of the problem situation (cf., Hallsworth, 2023). As such, they should keep in mind that they might learn something new from the people they are attempting to nudge and consider the use of methods that enable them to recognise if their understanding of the decision environment is mistaken.

Conclusion

Developing nudges through the lens of representational alignment can help with anticipating what kind of nudges could work and for whom, and when and why they might not work or stop working. This is in stark contrast with the cognitive miser view's blanket assumption of high representational alignment, which cannot account for these details and needs to explain discrepancies away by disparaging people's cognitive capacities. An increased focus on balancing the agency between people and choice architects – an acknowledgement that people think – may bring us closer to achieving both theoretical and practical progress.

References

Bargh, J. (2017). *Before you know it: The unconscious reasons we do what we do*. Simon and Schuster.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., & Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, *28*(8), 1041-1055. https://doi.org/10.1177/0956797617702501

Bergquist, M., Thiel, M., Goldberg, M. H., & van der Linden, S. (2023). Field interventions for climate change mitigation behaviors: A second-order meta-analysis. *Proceedings of the National Academy of Sciences*, *120*(13), e2214851120. https://doi.org/10.1073/pnas.2214851120

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980-989. https://doi.org/10.1038/s41562-021-01143-3

Bucher, T., Collins, C., Rollo, M. E., McCaffrey, T. A., De Vlieger, N., Van der Bend, D., Truby, H., & Perez-Cueto, F. J. (2016). Nudging consumers towards healthier choices: a systematic review of positional influences on food choice. *British Journal of Nutrition, 115*(12), 2252-2263. https://doi.org/10.1017/S0007114516001653

Chater, N. (2018). Is the Type 1/Type 2 distinction important for behavioral policy? *Trends in Cognitive Sciences*, *22*(5), 369-371. https://doi.org/10.1016/j.tics.2018.02.007

Chater, N., & Loewenstein, G. (2022). The i-frame and the s-frame: How focusing on individual level solutions has led behavioral public policy astray. *Behavioural & Brain Sciences,* 1-60. https://doi.org/10.1017/S0140525X22002023 de Ridder, D., Kroese, F., & van Gestel, L. (2022). Nudgeability: Mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science, 17*(2), 346-359. https://doi.org/10.1177/1745691621995183

Deetlefs, J., Bateman, H., Dobrescu, L.I., Newell, B.R., Ortmann, A.& Thorp, S. (2019). Engagement with retirement savings: it's a matter of trust. *Journal of Consumer Affairs, 53,* 917-945. https://doi.org/10.1111/joca.12208

Dolan, P., Hallsworth, M., Halpern, D., King, D., & Vlaev, I. (2010). *MINDSPACE: influencing behaviour for public policy.* London: Cabinet Office. http://www.instituteforgovernment.org.uk/sites/default/files/publications/MINDSPACE.pdf

Duckworth, A. L., Milkman, K. L., & Laibson, D. (2018). Beyond willpower: Strategies for reducing failures of self-control. *Psychological Science in the Public Interest*, *19*(3), 102-129. https://doi.org/10.1177/1529100618821893

Embrey, J. R., Donkin, C., & Newell, B. R. (2023). Is all mental effort equal? The role of cognitive demand-type on effort avoidance. *Cognition, 236*, 105440. https://doi.org/10.1016/j.cognition.2023.105440

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3)*,* 223-241. https://doi.org/10.1177/1745691612460685

Goldin, J., Homonoff, T., Patterson, R., & Skimmyhorn, W. (2020). How much to save? Decision costs and retirement plan participation. *Journal of Public Economics, 191,* 104247. https://doi.org/10.1016/j.jpubeco.2020.104247

Hallsworth, M. (2022). Making Sense of the “Do Nudges Work?” Debate. *Behavioral Scientist*, 2, https://behavioralscientist.org/making-sense-of-the-do-nudges-work-debate/

Hallsworth, M. (2023). A manifesto for applying behavioural science. *Nature Human Behaviour, 7*(3), 310-322. https://doi.org/10.1038/s41562-023-01555-3

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, 14-31. https://doi.org/10.1016/j.jpubeco.2017.02.003

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973-986. https://doi.org/10.1177/1745691617702496

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences*, *22*(4), 337-349. https://doi.org/10.1016/j.tics.2018.01.007

Jachimowicz, J. M., Duncan, S., Weber, E. U., & Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, *3*(2), 159-186. https://doi.org/10.1017/bpp.2018.43

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus & Giroux.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(1), 1-17. https://doi.org/10.1017/S0140525X00041157

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. https://doi.org/10.1353/bsp.2017.0010

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J., & Wagenmakers, E. J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences, 119*(31), e2200300119. https://doi.org/10.1073/pnas.2200300119

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. http://dx.doi.org/10.1561/105.00000091

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, *119*(1), e2107346118. https://doi.org/10.1073/pnas.2107346118

Michie, S., Van Stralen, M. M., & West, R. (2011). The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*(1), 1-12. https://doi.org/10.1186/1748-5908-6-42

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., ... & Duckworth, A. L. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, *600*(7889), 478-483. https://doi.org/10.1038/s41586-021-04128-4

National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care.* Washington, DC: The National Academies Press.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1-19. https://doi.org/10.1017/S0140525X12003214

Newell, B. R., & Shanks, D. R. (2023). *Open Minded: Searching for Truth about the Unconscious Mind.* MIT Press.

Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E-J., & van Rijn, H. (2015). On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making, 10,* 1-17. https://doi.org/10.1017/S1930297500003144

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Osman, M., McLachlan, S., Fenton, N., Neil, M., Löfstedt, R., & Meder, B. (2020). Learning from behavioural changes that fail. *Trends in Cognitive Sciences*, *24*(12), 969-980. https://doi.org/10.1016/j.tics.2020.09.009

O'Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., ... & Zrubka, M. (2018). Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science*, *13*(2), 268-294. https://doi.org/10.1177/1745691618755704

Ritchie, S. (2020). *Science fictions: How fraud, bias, negligence, and hype undermine the search for truth*. Metropolitan Books.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*(4), e56515. https://doi.org/10.1371/journal.pone.0056515

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., Tamman, A. J. F., & Puhlmann, L. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General*, *144*(6), e142 –e158. https://doi.org/10.1037/xge0000116

Sher, S., & McKenzie, C. R. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467-494. https://doi.org/10.1016/j.cognition.2005.11.001

Sher, S., McKenzie, C. R., Müller-Trede, J., & Leong, L. (2022). Rational Choice in Context. *Current Directions in Psychological Science*, *31*(6), 518-525. https://doi.org/10.1177/09637214221120387

Simmons, J., Nelson, L., & Simonsohn, U. (2022). *Meaningless Means #1: The Average Effect of Nudging Is d = .43.* http://datacolada.org/105

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), e2200732119. https://doi.org/10.1073/pnas.2200732119

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., & Aczel, B. (2018). A systematic scoping review of the choice architecture movement: Toward understanding when and why nudges work. *Journal of Behavioral Decision Making, 31*(3), 355-366. https://doi.org/10.1002/bdm.2035

Szollosi, A., Donkin, C., & Newell, B. R. (2023). Toward nonprobabilistic explanations of learning and decision-making. *Psychological Review, 130*(2), 546–568. https://doi.org/10.1037/rev0000355

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008-1018. https://doi.org/10.1016/j.tics.2020.09.005

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Thomson, K. S., & Oppenheimer, D. M. (2022). The “Effort Elephant” in the Room: What Is Effort, Anyway?. *Perspectives on Psychological Science*, *17*(6), 1633-1652. https://doi.org/10.1177/17456916211064896

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., ... & Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460-471. https://doi.org/10.1038/s41562-020-0884-z

World Bank Group. (2015). *World Development Report 2015: Mind, Society, and Behavior* https://openknowledge.worldbank.org/handle/10986/20597