Behavioural Insights and Public Policy:

The Evolution and Limits of Nudge Governance

Editors: Firas Abdul Jabbar, Zac Robb

Writers: Aarushi Sharma, Amane Tanaka, Anna Polubutkina, Ella Brittan, Kayane El-Ayache, Yi Swee Tan

The Wilberforce Society

Cambridge, UK

www.thewilberforcesociety.co.uk

October 2025

Disclaimer:

The Wilberforce Society is an independent, student-run think tank and a registered charity (no. 1169450). It is independent of the University of Cambridge and does not speak on its behalf. The views expressed are those of the author(s), not those of the University or of the Society, its trustees or membership at large. The Society is non-partisan and supports no political party. Policy proposals are made only to further its charitable objects and to inform public debate.

TABLE OF CONTENTS

I. Introduction

Fifteen years since the establishment of the UK's Behavioural Insights Team (BIT), the institutionalisation of behavioural science in government represents an evolution in public policy. A major component of BIT's work involves the concept of 'nudges'. These are subtle alterations to a person's choice environment that guide their behaviour without restricting choice (Thaler & Sunstein, 2008). Inspired by foundational research on heuristics, biases and choice architecture (Kahneman & Tversky, 1974), the BIT pioneered a model of 'test, learn, adapt' (Haynes et al., 2012) which has led to interventions with successes in domains ranging from vaccine uptake to organ donation. The appeal of this new form of behavioural governance is exemplified by the global influence and expansion of 'nudge units'.

The maturation of this relatively new field, and the BIT's evolution from a small, seven-person team to a global entity, raises crucial questions about the limits and scope of nudge, and behavioural science at large. Research involving Randomised Controlled Trials (RCTs) shows that nudge can be effective in targeted interventions, such as increasing police diversity (BIT, 2017) and improving tax compliance (BIT, 2012). However, persistent challenges remain across interventions. Some, such as long-term efficacy and scalability, are common to most policy interventions. Others are more acute for nudges. First, ethical concerns about covert paternalism: unlike traditional regulations that openly restrict choice, nudges are more vulnerable to accusations of influencing behaviour without transparency. Secondly, the risk of substitution: nudges can be cheap and politically palatable, with the potential to be deployed as an alternative to the more expensive or complex structural reforms. Given significant evidence to the contrary, the central policy question is not whether behavioural insights are useful, but how they can most effectively be institutionalised to improve outcomes over the long term.

This paper provides a comprehensive evaluation of the nudge unit model through the lens of the BIT, examining both domestic and international impact. Synthesising these insights, we aim to suggest a pathway for moving beyond initial, project-based methods of behavioural policymaking, and towards a robust model embedding empirical results within government and policy. Collectively, our recommendations outline a framework to harness the power of behavioural evidence not as an ad-hoc approach but as an integrated and learning-oriented component of policy infrastructure.

Introduction: Rational Actor to Bounded Rationality

From its inception in 2010, the United Kingdom's Behavioural Insights Team (BIT) marked a departure from the assumptions embedded in previous policy models. Models following the Homo economicus paradigm were proposed under the assumption of rationality; individuals had stable preferences, and would act to maximise utility when in possession of all the relevant information necessary to make a decision (Zey, 2001). This provided a framework for the formulation of social policies focused on economic incentives and provision of information

(Thaler & Sunstein, 2008). However, individuals commonly display predictable biases due to, for example, limited mental resources (Kuehnhanns, 2018) and cannot always be assumed to act in their own 'best interest'. This has historically led to persistent gaps between intended and observed policy outcomes, indicating a need for an alternative approach involving the application of behavioural research to policy-making.

The foundations for such an integration of cognitive science and behavioural economics can be seen in Kahneman and Tversky's work on heuristics and biases (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974). People tend to not behave in a rational manner, and their behaviour is constrained by limitations in mental resources, such as attention and memory. Human judgement is therefore often shaped by systematic, predictable mental shortcuts, or 'heuristics'. Knowledge of these shortcuts led to the development of dual-process theory (Kahneman, 2011), which posits that human cognition can be broadly divided into two types: intuitive, automatic 'System 1' thinking and conscious, analytical 'System 2' thinking. The shortcuts and biases of System 1, optimised for efficiency and preservation of cognitive resources, govern most daily decisions (Mullainathan & Shafir, 2013). The creation of the BIT initiated the embedding of behavioural testing in routine policy design. This was a novel institutionalisation of these scientific ideas, recognising the need for behavioural evidence-based governance and a consideration of citizens' bounded rationality (Simon, 1955).

In conjunction with dual-process theory, the then-emerging concept of libertarian paternalism (Thaler & Sunstein, 2003, 2008) in governance provided both the justification and philosophical framework for these insights to be translated into policy. Thaler and Sunstein argued for the use of 'nudges', which subtly influence choice architecture, choices that policymakers judge to improve welfare, while preserving choice. By not restricting freedom of choice through mandates or heavy-handed legislation, this framework resolved a key issue for liberal governments. The formation of the BIT therefore showed a commitment to this new form of behavioural governance, seeking evidence-based, cost-efficient solutions to policy issues based on a 'test, learn, adapt' philosophy (Haynes et al., 2012; Halpern, 2015).

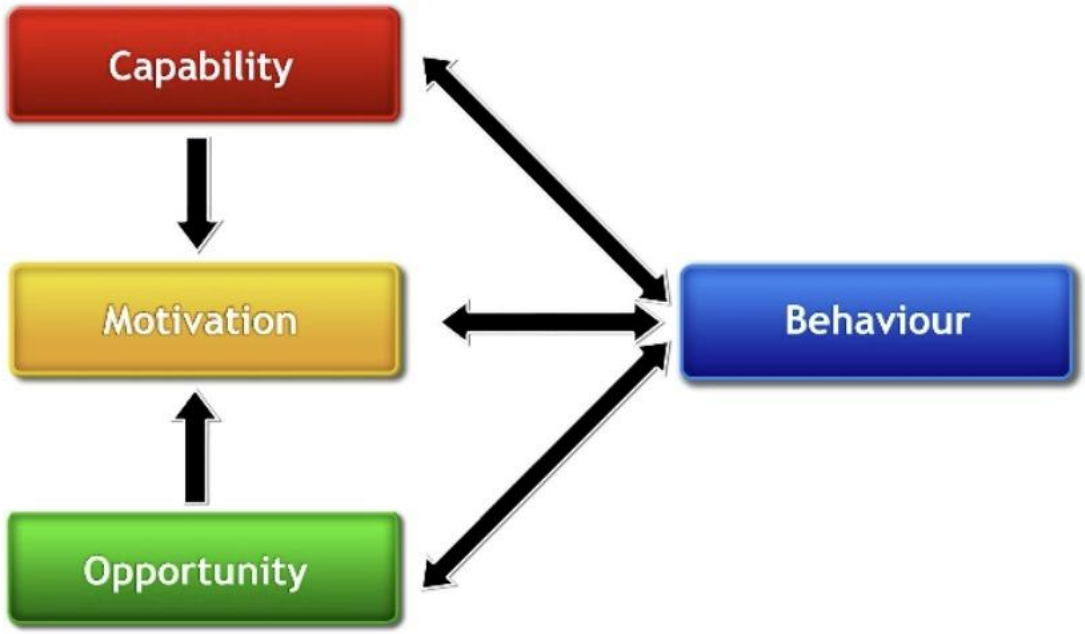
II. Theoretical Background

Definition of Nudges and Framework

A nudge is a change to the choice environment, such that human behaviour is influenced in a predictable way. Adopting the typical "what it is not" framing, we may also define nudges negatively. Nudges are not bans or mandates. They do not change incentives significantly, nor do they remove options. Nudges are also typically distinguished from training programmes or broad education campaigns, although this boundary is not always sharp, and some authors have discussed the possibility of 'educative nudges'. Nudges are a light-touch intervention, steering without coercing behaviour by redesigning the decision context.

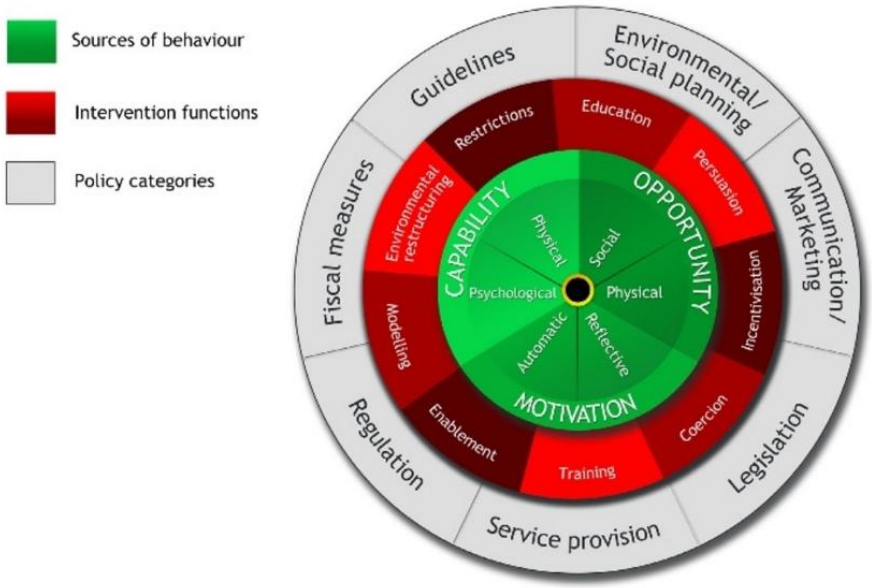
But before we are able to coherently discuss "nudge" policies, we need to be able to track how it works, that is, its mechanism of action. We should also attempt to model how nudge policies are developed.

Nudges are best understood as part of a broader toolkit of behaviour change interventions. Accordingly, we first develop a general framework for behavioural interventions, before zooming in on nudges. This will provide a basis for subsequent analysis of how nudge policies affect outcomes. Michie et al. (2011) proposes a three-layered Behavioural Change Wheel (BCW) framework for interpreting the impact of policy on behaviour.



The innermost layer of the framework is a behavioural system. In this model of behaviour, capability, opportunity and motivation interact with one another, and jointly produce a behavioural response. The behaviour in turn influences these three components. Hence, the framework is non-linear, in that components within a layer interact with each other. It is useful to think of interventions as acting upon one or more components in this system of behaviour.

With capability, a distinction is made between physical and psychological capability. With opportunity, the authors distinguish between physical and social opportunity. Motivation is divided between reflective (involving more planning and deliberation) processes and automatic (involving more impulses and emotions from innate dispositions) processes. These capture the range of behavioural mechanisms that may be involved in producing change. Modelling behaviour is important as behavioural interventions can sometimes be designed without having first established a theoretically-predicted mechanism of action, and instead be based on "implicit commonsense models of behaviour". This has the problem of excluding potentially important variables.



The middle layer consists of 9 classes of interventions, which are Education, Persuasion, Incentivisation, Coercion, Training, Restriction, Environmental Restructuring, Modelling and Enablement.

The outermost layer is composed of 7 categories of policies, which are Fiscal, Communication / Marketing, Service Provision, Legislation, Regulation, Guidelines and Environmental / Social Planning.

For instance, part of the English 2010 Tobacco Control Strategy involved removing tobacco products from display in shops. This is classified as a 'Legislation'-type policy and an 'Environmental Restructuring'-type intervention.

In terms of intervention development, the following steps are taken. After a target group and target behaviour are identified, the COM-B target(s) are diagnosed. The COM-B targets are then mapped to intervention types, which are in turn mapped to policy options. Pilot implementation typically acts as the vehicle for testing the efficacy of the intervention. During the implementation phase, testing and evaluation run concurrently, through A/B testing and constant outcome monitoring, among other methods. This helps, for instance, in gathering more robust evidence for causality, in fine-tuning nudges to account for heterogeneity within the target group, and in

knowing if the effect of the intervention persists over time, instead of only being effective on impact.

Background about Nudges

Next, one might think about how nudges have rapidly shot to prominence in recent years. Hence, we will introduce the context from which nudges have gained traction by tracing its foundations, which span the intellectual, ethical and practical domains. This section is crucial as key concepts are embedded within. These concepts are necessary for understanding the remainder of this paper and are woven into the structure of this section.

Intellectual Foundations

The field as an academic subject began with the pioneering research of Daniel Kahneman and Amos Tversky at the Hebrew University of Jerusalem in the 1970s. They were among the first researchers to test the Homo Economicus assumption experimentally, and kickstarted a "heuristics-and-biases program" of academic inquiry. Their contributions included seminal works such as "Judgment under Uncertainty: Heuristics and Biases" (1974), "Prospect Theory: An Analysis of Decision under Risk" (1979), "The Framing of Decisions and the Psychology of Choice" (1981), "Choices, Values, and Frames" (1984) and "Advances in Prospect Theory: Cumulative Representation of Uncertainty" (1992). These works claimed to demonstrate various ways in which human behaviour deviated from the classical "rationality" characterised inter alia by strict logical consistency and perfect Bayesian updating. Importantly, they argued not only that human behaviour deviated from this interpretation of rationality, but that it deviated systematically and predictably. Furthermore, they proposed a cogent framework to account for their wealth of novel findings, introducing concepts such as bounded rationality, heuristics and cognitive biases.

This extensive research culminated in the 2002 Nobel Memorial Prize in Economic Sciences, awarded to Daniel Kahneman and Vernon Smith "for having integrated insights from psychological research into economic science". It was a watershed for behavioural economics.

A crucial theoretical bedrock for nudges is the dual-systems model of cognition, popularised by Kahneman (2011). His distinction between "System 1" and "System 2" is a conceptual framework for thinking about how humans make decisions.

System 1 is engaged for mental activity that is rapid, automatic and intuitive. It produces reactions with little deliberate effort. It is involved in activities such as orienting to a loud sound or completing the phrase "bread and...". However, System 1's efficiency derives from its use of cognitive heuristics, which result in predictable biases.

System 2 gets activated for more complex tasks, requiring more reflective, and effortful cognition. Such tasks involve sustained attention, formal problem-solving, inhibition of impulsive responses and manipulation of working memory. An example is when one has to compare two washing

machines for overall value. System 2 is metabolically and psychologically "costly" to operate, so it tends to be selectively engaged rather than continuously active.

In sum, the central claim of the Behavioural Economics program is not that humans sometimes commit errors, but that erroneous biases are hardwired, in that virtually everyone errs in the same way. Like visual illusions, cognitive illusions are inevitable and persistent.

Nudges are based on the claim that many of our individual decisions are System 1-driven, with System 2 often left disengaged. Nudges are designed in such a way as to incline System 1 to select the option that System 2 would have endorsed had there been more deliberation.

Ethical Foundations

The ethical foundations of nudges rest on two core ideas. First, that libertarian paternalism can be an acceptable stance for policymaking. Second, that choice architecture is unavoidable.

Libertarian Paternalism

Thaler and Sunstein (2003) coined the term "libertarian paternalism" (sometimes described as soft paternalism) to bring together two notions which are typically held as incompatible. On the one hand, this approach is paternalistic in the sense that it involves government influence on people's decisions. Such intervening actions are justified in terms of citizens' supposed interest.

On the other hand, this policy stance is libertarian because it maintains that people do not lose any freedom of choice. The same set of options remains, and there is no explicit coercion.

The canonical analogy for nudges' place in individuals' decisions is the Global Positioning System (GPS) device (Sunstein 2015). This serves as a useful way to think about how nudges may comport with the idea of "libertarian paternalism". A GPS tends to "know" the entire network of roads (static data) and live traffic conditions (dynamic data) better than a human driver. It thus recommends a route, based on its own evaluation criteria and weights. The "best" path (as defined by the system) is then displayed in a salient manner, but the GPS does not sanction the driver for not adhering to it. The driver can deviate from the suggested path at any point, if for instance, a more scenic or emotionally familiar route is preferred. Nudges aim to function similarly, aiming to guide without forcing. Going further, some proponents may even argue that behavioural interventions enhance personal agency as it equips the individual with a more complete set of relevant information (Sunstein 2015).

Nonetheless, there is one crucial area where a GPS may differ from a nudge. A GPS may be called upon on demand, while some forms of nudges are, by design, unavoidable.

Choice Architecture

The justification for "libertarian paternalism" is closely tied to the concept of "choice architecture". Thaler and Sunstein (2008) invented the term "choice architecture" to describe

the way decisions are structured and presented—the design of the environment in which people choose. This includes the ordering of options, default settings and the framing of information. Importantly, their argument is that some form of choice architecture is always present. As Sunstein (2015) describes strikingly:

"Any store has a design; some products are seen first, and others are not. Any menu places the options at various locations. Any website has a design, which will affect what and whether people will choose. Streets, street signs, computers, cell phones, and ballots offer choice architecture of their own. Television stations come with different numbers, and strikingly, lower numbers are better, even when the costs of switching are vanishingly low; people are more likely to choose a station numbered 2 or 3 than one numbered 150 or 200."

Because a set of choices must necessarily exist in some context, choice architecture can never be neutral. This is an important recognition with regard to ethical implications. If choice architecture is unavoidable, then the relevant question ceases to become one of whether policymakers should influence decisions at all, but whether they should do so responsibly and transparently. In this view, nudges may represent less of an intrusion by government into personal choice than critics may initially assume. Rather than introducing influence where none existed, nudges operate within a space that was already structurally biased.

These first two foundations demonstrate that nudges have both psychological (positive) and ethical (normative) dimensions. For nudges to be credible candidates for public policy, they must be theoretically cogent and morally palatable. These are necessary conditions for potential government adoption.

Practical and Political Foundations

The financial success of many early nudge programs is notable. Nudge policies are extremely low-cost by design, usually of the order of a few thousand pounds, like a typical monthly salary. For example, a campaign by the White House Social and Behavioural Sciences Team (SBST)—the US nudge unit—to increase savings among military personnel cost only $5000.

This is because nudges usually involve relatively lightweight adjustments, such as introducing an additional message in a letter. In addition, nudges often leverage on infrastructure that is well-established, such as the existing communications network. This was the case for the savings campaign example, which harnessed the existing email system.

Behavioural effects are frequently modest in percentage terms, thus causing the impact of a nudge-driven campaign to appear modest. However, small proportional effects can translate into large absolute gains when applied at scale. Using the same example, the treatment group had an enrollment rate ranging from 1.6 to 2.1%, compared to the 1.1% rate for the control group. But in absolute terms, the campaign was estimated to have generated $1,600 in additional savings per dollar spent by the government, which represents an impressive impact to cost ratio.

Finally, low cost has strategic significance for policymaking. Since the cost of behavioural interventions are small, the break-even threshold is correspondingly low. Thus, the low-cost profile of nudge policies make it a low-risk endeavour, reducing political barriers to implementation. In the context of the UK, its nudge unit was established in 2010 by the 2010–15 coalition government, which ushered in a period of austerity in the wake of the 2008 financial crisis. Its nudge campaigns have achieved striking financial returns (in percentage terms at least), making them very attractive to implement and sustain over a long period.

The UK also has the unique experience of instituting the world's first nudge unit. This has served as a platform for the BIT to impart their expertise to other countries, deepening international ties.

Motivation for Applying Nudges

Systematic deviations from rational-choice benchmarks can be policy-relevant because they suggest that, in some contexts, individuals may fail to act in ways that advance their own long-run welfare, although what counts as a person's 'best interests' is itself contestable. This can result in persistent welfare losses. And because these losses are not random, but arise predictably, policy can address them through targeted changes to the choice architecture. Seemingly minor and specific design details of policy—such as framing, defaults, friction—matter for outcomes, and can be targeted to influence behaviour, thus creating scope for low-cost interventions.

Classification of Types of Nudges

This section aims to summarise the typical nudges, hence giving form to subsequent objects of analysis.

Behavioural Mechanisms

We will first outline the main mechanisms that nudge the target.

Inertia/status quo. People tend to stick with the existing option or the option they are passively placed into, even when switching would benefit them.

Example: Automatic enrolment into a workplace pension with an easy opt-out.

Limited attention/salience. People focus on what's most visible or top-of-mind and often overlook information that's buried or complex. Thus decisions can be driven by what is immediately salient rather than what matters most.

Example: Placing healthier food at eye level or near the checkout counter significantly increases selection.

Present bias/procrastination. Future benefits are discounted heavily against immediate pains, so people delay actions they would have endorsed with System 2.

Example: Commitment devices to lock in future behavior, such as to encourage higher retirement savings.

Friction/hassle costs. Small practical barriers (extra steps, time, paperwork) can sharply reduce take-up. The reduction is disproportionate to the small costs.

Example: Pre-filled forms for benefits.

Social norms. People are influenced by what they believe others do or approve of, and will often conform to perceived norms.

Example: Energy reports that compare a household's usage to similar neighbours and highlight "efficient" households.

Framing/reference dependence. Choices depend on how options are presented and what the reference point is. Two logically equivalent framings can yield different outcomes.

Example: Framing in terms of a loss rather than a gain.

Complexity/choice overload. Too many options or complicated information can lead to confusion, decision avoidance, or reliance on crude heuristics.

Example: Tiered menus for public services, with a small number of tiers.

Examples of Nudge Policies

Policy	Description	Examples
Default options (automatic enrollment)	Shifting the recommended option as the pre-selected choice	Opt-in vs opt-out by default: Organ donation, savings programs, free school meals programs; Pre-scheduling vaccination appointments by default
Information about peers (social norms feedback)	Provide people with information about what similar others do	GP antibiotic prescriptions, household energy use, hotel towel reuse
Manipulating friction	Increasing or decreasing friction by changing the time or cognitive effort required to complete an action	Reducing friction: Pre-filled forms, 'one-click' navigation pathways for vaccination scheduling. Increasing friction: Extra confirmation steps for antibiotic prescriptions, cooling-off durations

Framing	Change how options are presented, while keeping the set of options unchanged	Positive vs negative framing; category framing; reference-point framing; rearranging items in a grocery store
Reminders / Prompts	Sending a short, simple and action-oriented cue, timed close to the point of taking action	Public services (taxes, bills, appointments, licences)
Provision of information	Provide clear, relevant information when mistakes are likely due to lack of awareness or difficulty comparing options	Leaflets about the benefits of climbing stairs; information on stickers displaying the full costs of durable goods over the lifecycle; intuitive traffic light labels to indicate foods' health impacts

Taxonomies of Nudges

Two main taxonomies are commonly used to categorise nudges. One approach groups nudges into three broad buckets: Structure, Assistance, and Information. Structure nudges target the organisation of choice alternatives—how options are presented. They are typically effective because decisions are shaped by inertia, as well as by limited attention and cognitive overload. Examples include defaults, option ordering and prominence, simplification (e.g.\ streamlining forms), and partitioning (e.g.\ how alternatives are grouped).

Assistance nudges reinforce behavioural intentions, helping individuals follow through from intention into action. They target common implementation failures such as present bias, limited self-control, procrastination, and forgetfulness. Examples include reminders and prompts, planning aids, commitment devices, and feedback mechanisms such as progress tracking.

Information nudges focus on the description of alternatives, changing what people notice, and thus how they interpret a decision. This type of nudge is motivated by the notion that choices often rely on heuristics and are sensitive to salience, framing, and misperception. Examples include framing and reference points, salient disclosure and labels, and social norms information (e.g.\ comparisons or norms-based feedback).

A second cross-cutting way to categorise nudges is to distinguish between Type I and Type II nudges, corresponding to the 'System' they act upon. A Type I nudge is designed to influence behaviour without engaging System 2, leveraging humans' supposed in-built cognitive bias. A Type II nudge is intended to trigger the deliberative processes of System 2, by promoting a sustained reevaluation of the evidence base on which a choice is made.

For example, in the context of food portion control, a Type I nudge is to reduce plate sizes at food establishments, while a Type II nudge is to provide calorie counts on food menus.

Behavioural Economics vs Economics (Homo Economicus)

At this juncture, the relationship of behavioural economics with "standard" economics may lurk in the mind of the reader. How do these two domains relate to each other? Are they fundamentally at odds?

Behavioural economics has been sometimes presented to audiences as an obvious response to the gaping oversights contained within the absurd assumptions of "standard" economics. Sometimes, behavioural economics is even regarded to be able to sweep away vast areas within the compendium of economics, by overthrowing its brittle foundations. Is behavioural economics the critical blow to "standard" economics it is sometimes made out to be? Should Homo Economicus be everywhere replaced by Homer Simpson? Although these notions have gained some traction, they are misguided.

Like any other formal, rigorous and mathematical discipline, economics is frank about its unrealistic assumptions. But assumptions are an essential feature, not a shortcoming. The parable "On Exactitude in Science" describes a cartographer who created a map on a 1:1 scale. This life-sized map was so perfect that it was rendered useless, as it would block out the sun, and the country itself could be used as its own map. This is a cautionary tale for how perfect fidelity makes a model redundant. Abstraction and simplification are necessary. At different "levels" of analysis, it is perfectly sensible to model an object to different degrees of granularity. An obsession with capturing every detail and oddity makes a theory a burden rather than a tool.

Much of science adheres to a workflow as such. First, a complicated system is simplified by assuming away some of the complications. This strategic omission allows us to specify a smaller number of parameters. Next, one derives the implications of the specified system, obtaining a benchmark result from this idealised case. Finally, the assumptions are stripped away in turn. The complications return to the model, and one computes how each complication affects the idealised result. In the context of economics, behavioural economics may be thought of as one of the components in this final stage. The two fields are fundamentally concerned with answering different sets of questions. They have a relationship characterised more by complementarity than by opposition.

The benefits to this approach are clear. First, it permits superior tractability. The problem becomes easier to conceive in a human mind, and a solution is more easily obtained. Moreover, benchmark results generated under idealised conditions often provide illuminating insights to the mechanisms behind complex phenomena, which may have otherwise been obscured in the noise had the problem not been simplified.

More generally, this approach exhibits the power of a theory—the ability to derive widely applicable implications from a set of few starting assumptions. In many contexts, Homo Economicus is a powerful analytical tool.

In sum, economics makes no pretense about the mutability of its approach to modelling agents. Assumptions are always subject to refinement or even outright repudiation. As an empirical field, economics by its nature always stands open to changing beliefs about how the world actually works.

Criticisms of Behavioural Economics

Despite the high-profile recognition the field has received, behavioural economics remains a controversial area. Before assessing either the efficacy or the ethical permissibility of nudge policies, it is important to engage with several major objections to the underlying theory. This section offers a brief critical examination of the intellectual foundations of nudging, outlining the principal charges levelled against it, and notable debates within the field.

A major criticism of behavioural economics is that its experiments lack external validity. Much of the empirical evidence (including that of Kahneman and Tversky) originates from highly contrived experimental environments. For instance, experiments that attempt to demonstrate the significance of "Framing" effects often involve hypothetical choices. Results may also be mere artefacts of other conditions that are peculiar to the lab, such as the pressure of being observed and the unfamiliarity of the situation. Human behaviour in such settings may not generalise. This highlights a fundamental drawback of lab-based experimental psychology. Research by other experimental economists suggest that humans make better decisions and approach neoclassical rationality if given time and the opportunity to learn from experience, which belies the central tenet (and central justification for application to policy) that humans deviate from rationality in a systematic manner.

Some critics argue that several heuristics remained underspecified for a long time relative to formal economic models, making them vulnerable to post hoc explanation, fitting the data after the fact. However, later work, including resource-rational approaches, has sought to place bounded cognition on firmer formal foundations.

There have been instances where key findings in behavioural research were found to be specious. One example is priming. Priming refers to the incidental influence of environmental context on cognition and behaviour and is a cornerstone of behavioural science theory and its practical applications. However, it has been subject to replication crises. In 2012, just a year after the publication of the book *Thinking Fast and Slow*, the significance of priming effects was victim to an acute replication crisis. Kahneman published an open letter to researchers in the field of "social priming", acknowledging that the field was "the poster child for doubts about the integrity of psychological research." "Social priming" has also been subject to accusations of being ambiguously defined, and having no consensually agreed upon meaning. In a 2017 comment on a blog post, Kahneman himself admitted that he has "changed [his] views about the size of behavioural priming effects", and that he "placed too much faith in underpowered studies". If some influential findings are weaker than initially thought, this should make policymakers more cautious about how behavioural evidence is used. At the same time, priming was never central to most real-world behavioural public policy, so replication problems in that literature do not, on their own, invalidate behavioural interventions more generally.

III. Domestic Case Study: The BIT

Institutional Architecture of the BIT

The BIT's unique infrastructure, characterised by its evolution from a small, seven-person team within the Cabinet Office to a part-privatised company, has played a significant role in its efficacy as an institutional vehicle. Between 2014 and 2021, 'Behavioural Insights Ltd' was owned jointly by the UK government, the innovation agency Nesta, and its employees. This tight-knit, relatively non-hierarchal structure helped the BIT to mitigate common problems faced by teams operating within a traditional bureaucratic structure, including risk-aversion, departmental silos and short-term political cycles. As of 2021, the company is fully owned by Nesta. BIT also runs a permanent office in Singapore since 2016 and partners with many government ministries to work on projects related to innovative social policy.

This unique evolution helped the BIT introduce an experimental approach to governance in three ways. Firstly, the BIT emphasised methodological rigour as foundational to its function. While academics have long advocated for the importance of Randomised Controlled Trials (RCTs) in behavioural policy (e.g.\ Haynes et al., 2012), the BIT was instrumental in making them a routine component in its approach to policy development. Its unique institutional structure gave team members the ability to create and deploy trials, potentially at a faster rate than can be seen in traditional academia. This transformed the BIT into a 'skunkworks' for policy (John, 2014), wherein the experimental method was part of the 'core product'.

Second, it allowed the BIT to have financial and operational autonomy. Operating using a cost-recovery model, and reinvesting profits (Nesta, 2021) meant that the BIT could continue its UK government work while also contracting with other governments, local authorities, and international organisations. Thus, it escaped restrictive government pay scales and annual budget issues (Halpern, 2015). This relative freedom gave rise to an entrepreneurial culture focused on the impact of policy interventions, and creating a virtuous cycle: successful RCTs for new interventions create revenue and increase credibility, generating funds for further experimentation. As noted by John (2014) the BIT effectively functions as a 'policy entrepreneur' to promote innovation by navigating between market and state interests. However, it is worth considering that a focus on cost recovery could lead to the prioritisation of more marketable projects, moving focus away from pressing policy issues.

Third, as mentioned, the 'Ltd.' structure of the team allowed the BIT to collaborate directly with national and international partners. Interventions designed within the UK could then be refined, and exported internationally to enable testing in varied contexts. Such international applications, including work in Argentina and Australia (Brown et al., 2024; BIT, 2023), further validated the results of trials and allowed for more sophisticated approaches to be developed across cultures. The BIT's institutional structure was an essential aspect for the fulfillment of its 'test, learn, adapt' philosophy (Haynes et al., 2012), allowing it to function in its role as a generator of evidence informing behavioural public policy.

The EAST Framework

For the operationalisation of behavioural science to occur at scale, a translational framework that can be understood by policymakers without psychological training is required. The 'EAST' framework (Easy, Attractive, Social, Timely) serves this purpose by providing a format for intervention constructions and hypothesis generation (BIT, 2014). BIT has since refined its approach, and published a revised version of EAST (BIT, 2024) with increased contextual depth, taking into account a decade of RCT evidence. Where possible, concepts from the revised additions have been incorporated into the existing EAST principles discussed below. Each mechanism can be viewed as engaging with specific psychological principles to overcome barriers to inducing behavioural change:

Easy: This principle aims to reduce decision friction and cognitive load by applying behavioural insights on defaults (Johnson & Goldstein, 2003) where individuals display a tendency to stick to options which are pre-selected. This may be due to loss aversion, bias towards the status quo, or the additional mental effort required to change a decision, aligning with the goals of System 1 thinking (Kahneman, 2011). Recent work (BIT, 2024) has drawn attention to the reduction of 'sludge', highlighting that reducing administrative friction can act in conjunction with simplified choices. This principle can be leveraged by reducing steps, simplifying interventions wherever possible, or switching defaults.

Attractive: This principle targets motivational salience and attention. Popular theories of salience (Bordalo, Gennaioli, & Shleifer, 2012) posit that attributes that are novel, easily accessible or emotionally salient tend to capture our attention disproportionately. By using personalised language or bold formatting, nudge interventions can increase the salience of the desired choice. However, this principle can also be viewed through the lens of motivational crowding theory (Frey & Jegen, 2001) which holds that intrinsic motivation can sometimes be hindered by the promise of extrinsic rewards. Any interventions seeking to leverage this principle therefore require careful deliberation to effectively induce behavioural change while considering citizens who may already be predisposed to making the desired choice. The revised EAST framework discusses how effective incentive structures can be implemented, including gamification and prosocial incentives, where an option is incentivised by the fact that it benefits others.

Social: This principle considers citizens' fundamental need for social conformity using the social proof principle (Cialdini, 2007) which states that people determine the correct behavior in a given situation by observing what others do. Interventions involving this principle might use descriptive norms, describing the behaviour of others, or injunctive norms, describing the behaviours others approve of (Schultz et al., 2007). While often highly effective, interventions using this principle are context-dependent, as social norms can often reinforce negative behaviours if the norm being communicated is not the desired option.

Timely: This principle recognises that a person's willingness to act is state-dependent and may be transient. Hyperbolic discounting (Laibson, 1997) shows that individuals tend to prefer smaller, immediate rewards over larger, delayed ones. Therefore, an intervention that can break down long-term policy goals into tangible, short-term steps can help overcome this bias.

Furthermore, if interventions are timely people may be more likely to have the cognitive resources to engage with them, relating to the theoretical concept of cognitive scarcity (Mullainathan & Shafir, 2013).

Case Studies: Analysing Mechanisms and Context

The following section examines case studies from the BIT through the lens of the institutional structure and theoretical framework discussed above to determine which conditions are essential for success and failure in behavioural change. As will be discussed, the degree to which the mechanism used fits within the context of the intervention is particularly relevant.

Tax Compliance (2012)

Problem and Behavioural Diagnosis: The administrative letter received to seek late payment of income tax is a clear example of a friction point, where intention to pay does not translate into the described behaviour. It is possible that for at least a subset of citizens, persistent late payments stem from some form of present bias, where the immediate inconvenience of payment seems to outweigh the possible future punitive consequences of not paying, rather than an inability to pay. A study of 22,000 tax-filers (Martinez et al., 2022) supported this explanation. Substantial increases in filing probability can be observed close to tax deadline; a portion of this was attributed to present bias.

Intervention and Theoretical Mechanism: A simple descriptive norm was used to redesign HMRC tax letter reminders: "9 out of 10 people in the UK pay their tax on time", with some letters also containing information about local payment rates. This leveraged the social proof principle (Cialdini, 2007), increasing the salience of the desired behaviour being performed by the majority, activating a desire for conformity and reputational protection in the recipient. A clear and simple one-sentence cue was used to gain attention.

Result and Analysis: An RCT showed a statistically significant increase in payment rates, with a 15 percentage point increase between the control letter and the localised social norm letters. Analysing the contextual conditions of this trial can provide insight to the optimal conditions for nudges to be effective. Paying taxes is generally a behaviour that involves only a single action (as opposed to a sustained effort), so it can be inferred that the behavioural barrier is likely attentional rather than motivational. That is, most people intend to pay their taxes, but may not be aware of the social norm or pay enough attention. Therefore, this nudge intervention leveraging social norms may have been effective as it targeted this behavioural barrier in a context where the desired action was already the goal of most participants in the study. It can be seen from this that nudges are particularly effective for behavioural change when small effect sizes for one-off decisions can result in massive returns due to scale, as in the case of tax payments.

Organ Donation (2013)

Problem and Behavioural Diagnosis: Once again, a gap between action and intention is exemplified by low organ donor rates in the UK, despite high public approval. A consultation outcome under the 2016 to 2019 Conservative government suggested that while 80% of people said they would be happy to donate their organs after their death, only 37% were registered as donors (Department of Health and Social Care, 2020). Behavioural barriers in this case include omission bias (wherein negative consequences of inaction are perceived as less severe than those resulting from action), and possible status quo bias which is unavoidable within an opt-in organ donation system (Johnson & Goldstein, 2003).

Intervention and Theoretical Mechanism: The BIT tested a variety of messages, including a reciprocity appeal "If you needed an organ transplant, would you have one? If so, please help others" (BIT, 2013). This intervention targeted the conscious and thoughtful reflection of System 2 thinking (Kahneman, 2011) and activating the social norm of reciprocal altruism (Fehr & Gächter, 2000).

Result and Analysis: The reciprocity intervention led to 1,171 more sign-ups than the control prompt alone; a successful demonstration of the nudge approach. Although messaging framed around loss ("Three people die every day because there are not enough organs.") led to comparable initial click-through rates, those who saw the reciprocity message were significantly more likely to complete the registration process once clicking, making it the preferred choice for national rollout. In May 2020, the UK moved to an 'opt-out' system in England, an example of default switching. This highlights a hierarchy that is critical in behavioural policymaking. Although nudge interventions may be effective within an existing choice architecture, more overt changes to the architecture, such as in the default rule, can often achieve greater impact. The value of nudges (in addition to their main effect) can be in highlighting existing choice architecture that may not be optimal, and provide empirical evidence to justify broader structural reforms.

Increasing Police Diversity (2017)

Problem and Behavioural Diagnosis: Barriers contributing to the underrepresentation of ethnic minorities in UK police forces may be psychological, and not just due to a lack of interest. Stereotype threat, a situational problem in which individuals fear conforming to negative stereotypes about their own social group, can lead to anxiety and cause underperformance (Steele, 1997). This may have contributed to underrepresentation, along with a lack of belonging.

Intervention and Theoretical Mechanism: The BIT redesigned the application form to be simpler (targeting the Easy principle within the EAST framework), and removed any language that may have caused anxiety. The psychological principle of priming (where exposure to an initial stimulus influences your response to a subsequent stimulus) was used to reduce any anxiety and predispose participants towards good performance on a Situational Judgement Test (SJT) used as part of the recruitment process. This involved beginning emails with 'Congratulations!' and ending with 'Good luck!'. The form also asked participants to consider their motivation for

applying to be a police officer, an addition intended to prime applicants from underrepresented demographics to consider their presence within the police force as a valued identity for members of their community.

Result and Analysis: Non-white participants in the treatment group gained 12 percentage points in their percentile ranking on the SJT. In the field, the intervention increased a non-white applicant's chances of passing the SJT by 50% (BIT, 2017). This study demonstrates that nudges are useful for addressing performance gaps created by testing context, such as situational anxiety for marginalised groups. The nudge did not change applicants' skills or the content of the test. Rather, it altered the psychological environment to allow existing competencies to surface, improving procedural fairness without any changes to admission standards.

The insights derived from these case studies can be used to inform the future of behavioural government units. Additionally, nudge interventions might benefit from a tiered approach to change: subtle, low-cost nudges can be used to improve existing systems (as with HMRC), while the evidence generated from these nudges is used to drive broader, institutional reforms (as with organ donation). Finally, in addition to behavioural change, nudge units must be competent in their understanding of operational realities and the culture of the institution they are working with.

Portfolio Coding of BIT Interventions

While the preceding case studies offer insight into mechanism-context fit, they are not sufficient to answer the question: 'What tends to work in which context, and under which behavioural principle?'. To generate insights across the spectrum of BIT interventions in addition to individual case studies, an analysis of the BIT's published body of work has been undertaken. The aim of this section is to map patterns of success, comparing EAST principles with associated domains of policy. This portfolio coding is intended to be primarily descriptive rather than evaluative, facilitating hypothesis generation for future research over asserting causal claims. It should be noted that the insights derived from this analysis are exploratory in nature, and statistical representativeness should not be assumed.

The analysis draws on the BIT's own searchable database of academic publications, which can be filtered by policy area. The studies were classified into policy domains (Table 1) based on the BIT's own focus areas, which can be found in their project portfolio.

Studies from the BIT website were included if they:

- Were conducted by, or in collaboration with, the BIT
- Involved an RCT, quasi-experimental study (where an independent variable was manipulated without random allocation of subjects), or any controlled design comparing a dependent variable before and after an intervention
- Reported sufficient information for the outcome of the study to be determined
- Were available publicly

Qualitative studies, literature reviews, and any pilot studies without outcome data were excluded. A search for academic publications on the BIT website yielded 54 results. Of these, 36 fit the inclusion criteria and were used. Details of the excluded studies, and justification for why they were excluded, can be found in the appendix.

Coding Rules

Outcome:

- *Success:* Statistically significant positive effect on primary outcome ($p < .05$), no reported adverse effects
- *Mixed:* Significant effects only for subgroups, significant on secondary but not primary outcome, or one intervention being significant while the other was not
- *Null/Negative:* No significant effect

EAST Principle:

- *Easy:* Reducing friction, simplifying, defaults, reducing cognitive load
- *Attractive:* Getting attention, personalisation, incentives, salience
- *Social:* Norms, social proof, reciprocity, peer effects
- *Timely:* Prompting at right moment, planning prompts, reminders

In cases where the primary EAST principle was ambiguous, a secondary principle was added with this being noted in the analysis. The spread across time periods for studies was roughly even.

Of the 36 usable interventions, 27 were clear successes, 8 were mixed, and 1 was null or negative.

Analysis

Intervention outcomes by Policy Domain

Policy Domain	n	Success	Mixed	Null/Negative	Success Rate
Health + Public Health	6	6	0	0	100.0%
Social Capital and Prosocial Behaviour	5	3	2	0	60.0%
Tax and Revenue	7	6	0	1	85.7%
Education	6	5	1	0	83.3%

Government and Public Services	5	2	3	0	40.0%
Economy and Consumer Behaviour	5	4	1	0	80.0%
Environment	2	1	1	0	50.0%
Total	**36**	**27**	**8**	**1**	**75.0%**

Success rates varied considerably by domain. Health, Tax, Education, and Economy showed high success rates (80–100%), while Social Capital, Government, and Environment showed lower success rates (40–60%).

Intervention outcomes by primary EAST principle

Primary EAST Principle	n	Success	Mixed	Null/Negative	Success Rate
Easy	8	6	2	0	75.0%
Attractive	11	5	5	1	45.5%
Social	13	12	1	0	92.3%
Timely	4	4	0	0	100.0%
Total	**36**	**27**	**8**	**1**	**75.0%**

Timely interventions showed the highest and most consistent success rate (100%), though the small n (4) limits generalisability. Social interventions were most common and present in all domains except Government and Public Services, with strong overall success (92.3%). Easy interventions also had a fairly high success rate (75%). Attractive interventions showed the lowest success rates (45.5%) with an equal number of studies with successful and mixed results. This could hint at a higher susceptibility to backfire effects, but might also be due to a larger proportion of these interventions being implemented alongside a secondary EAST principle.

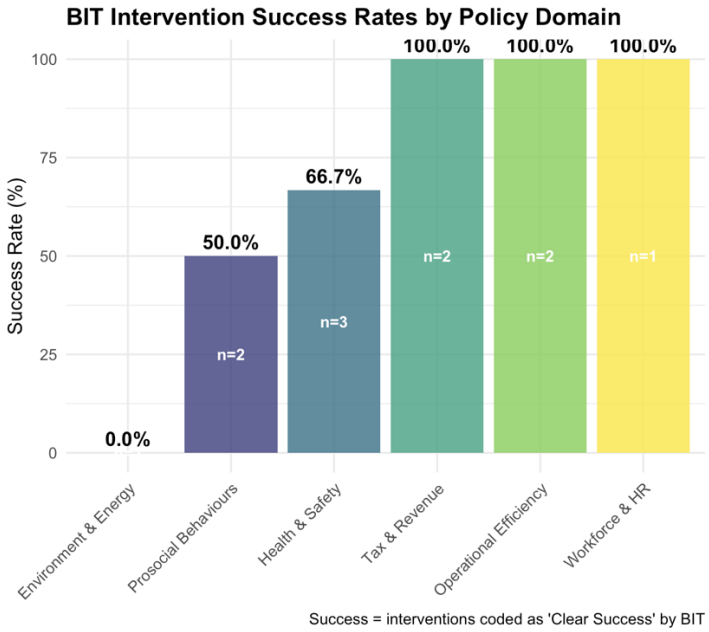


Figure 1: Success Rates by Domain and Primary EAST Principle

Exploratory Statistical Tests

A Fisher's exact test was conducted to examine the association between policy domain and study outcome (success vs.\ non-success). The results indicated no statistically significant association ($p = .299$). Despite the lack of statistical significance, notable variation in success rates was observed across domains, ranging from 40% in Government to 100% in Health. This pattern suggests that while domain may influence the effectiveness of behavioural interventions, the small sample size ($N = 36$) may lack sufficient power to detect such an effect.

A second Fisher's exact test examined the association between EAST principle and study outcome. The result was statistically significant ($p = .042$). Variation was observed across principles: Timely interventions achieved 100% success (4 of 4), Social interventions 92% (12 of 13), Easy interventions 75% (6 of 8), and Attractive interventions 46% (5 of 11). The lower success rate of Attractive interventions, combined with the success of Timely interventions, suggests that principle choice meaningfully influences intervention effectiveness. Nevertheless, the limited sample size warrants caution in interpreting these findings. Further research with larger samples would be necessary to draw definitive conclusions.

IV.Overseas Case Studies

Singapore

Overall History

One of few examples of how behavioural economics is widely used in a 'semi-authoritarian' country (a country with centralised governance and dominant-party system) is the various design thinking units that exist in Singapore. Singapore has informally applied behavioural economics long before establishing official dedicated units, with first such campaigns appearing as early as in the 1960s. One of the examples of earlier campaigns include the "Stop at two" population planning program, in which the Singaporean government strongly discouraged having more than two children through various financial incentives and public and social pressure. Although this campaign cannot be considered a direct use of nudges, it was nevertheless a behaviourally informed state campaign that marked the beginning of the use of behavioural insights in Singapore's public policy. Starting from the 2010s, Singapore's government has actively applied such behavioural insights in many policies by creating its own "nudge units" in various government departments, which marked a more systematic and evidence-based approach to using such insights. This includes separate units in the Ministry of the Environment, Prime Minister's Office, Ministry of Communication and several others.

Development of the Examined Unit

The most studied one appears to be "THE Lab" (renamed "Innovation Lab" in 2017), which was established in 2012 by the Prime Minister's office as one of the pioneering behavioural insights units in Singapore and which has been working for the benefit of the government ever since. It is important to note that although it is officially called a design lab, it uses many behavioural science insights in developing its interventions, which means it will be useful for our discussion. Innovation Lab is a part of PSD (Public Service Division), a unit that oversees HR for government officers across 16 ministries. The purpose of establishing this behavioural research group was to incorporate behavioural economics insights into the policies, service delivery, and operations of various departments. While being a relatively small unit with only six employees, it has been successful in collaborating with various government agencies with the main goal to redesign some existing policies so that they better correspond to the needs of the population.

Scope of Work and Methodology

As mentioned above, the main mission of the unit is to reframe policies to make Singapore's population better off. Innovation Lab works on providing cross-government digital services, financial products disclosure, and service design. The unit strives to inject human-centricity into public policies and services. Its main contribution towards the work of PSD lies in facilitating inter-government and government-citizen collaboration and co-creation. Among many different projects, the unit also works on empowering low-income households, encouraging residents to engage in activities that improve Singapore's living environment, and facilitating access to childcare facilities. It is important to note that the scope of the unit's work differs from traditional behavioural testing, and it is a different model of behavioural governance which is based on human-centred design and developing pilot projects.

Many of Singapore's behavioural insights units, including Innovation Lab, have successfully applied principles similar to the EAST framework designed by the UK's BIT. By making the designed interventions easy, attractive, social, and timely, Innovation Lab has developed many successful projects.

The unit also uses three core "innovation mindsets" that help to develop appropriate policies. This includes "empathising" (understanding all stakeholders' needs to better diagnose issues), "collaborating" (working across multiple agencies/departments and with all stakeholders for holistic and cohesive outcomes), and "experimenting" (trying ideas, testing assumptions and gaining evidence-based validation for existing proposals).

This "innovation mindsets" approach also indicates that the unit is more focused on design thinking than RCTs, and there is no explicit evidence in the unit's reports that RCTs are being used as a core method. However, as mentioned above, the unit still employs experimentation in a broader sense when designing policies, and RCTs may serve as one of the possible tools.

Example of a Completed Project

One of the most successful projects of Innovation Lab is arguably "Moments of Life". Started in 2018 as a relatively small project, its key mission was to help parents with young children to complete all administrative procedures, as parents of newborns in Singapore have to navigate multiple agencies and platforms in order to register births, apply for bonuses, check immunisation schedules, and complete other important tasks. This was especially challenging for parents who already have a lot of pressing obligations, and thus these complicated fragmented processes created confusion and friction among citizens.

The project is an example of how Innovation Lab uses its "innovation mindsets". The approach used to develop an appropriate response included completing ethnographic studies and in-depth interviews with parents of small children, which helped to identify key problems. The result of the project was an app that organised services for parents based on the current life moment of their child and provided single-entry access to multiple services, which significantly increased the rate of timely completion by parents of all important legal procedures required after the birth of a child. From a behavioural perspective, this shows that reducing administrative burdens and

providing timely reminders can help encourage parents to complete all important tasks, as the process now appears significantly easier, even if it is not actually very different from the original one.

Later, the project expanded to become a broader version of itself named "LifeSG", offering services that are not only in demand by the parents of newborns, but also necessary at various life stages: this includes employment, benefits for seniors, and some other services.

Limitations

The Innovation Lab has not disclosed any projects that could have been considered as failed ones. However, it is possible to say that the whole concept of operating as a "central consulting unit", which was present for several years after the unit was formalised in 2016, was not a good one as the Lab has later admitted itself. The problem was in doing projects for other agencies, which was perceived by these agencies as solutions suggested being "external" to them, which led to slower implementation and lower engagement if the concept proposed by Innovation Lab was good. Later, however, the unit shifted to another model, in which agencies form their own project teams, but the Lab instructs and consults these teams.

Another limitation is the lack of authority over other ministries and agencies. This suggests that the unit only works with other institutions when those institutions are willing to collaborate themselves; in other words, Innovation Lab cannot force collaboration even if the unit has a clear understanding of how a certain process can be improved inside an agency, but that agency is not willing to adopt it. This also demonstrates the importance of a clear governance structure that helped the UK's BIT and other nudge units around the world to become effective in developing and implementing their projects.

Australia

Overall History

Australia is also known as one of the early adopters of behavioural economics in policymaking. The Australian government has been exploring the use of behavioural sciences starting at least in 2007. The Australian network of behavioural insights teams is very broad, including units at central, departmental and regional levels. One of the first units in the country was established at a state rather than federal level. The NSW (New South Wales) Behavioural Insights Unit created in 2012 was largely supported by the BIT and focused primarily on conducting trials for various policies and regulatory measures. In 2015, before the creation of the federal institution, another regional nudge unit was established: Victoria Behavioural Insights Unit started working with policy design and testing.

Development of the Examined Unit

The creation of the federal behavioural insights unit, Behavioural Economics Team of the Australian Government (BETA), was first announced in 2015 by then-Prime Minister Malcolm Turnbull as a part of the government's innovation agenda. Experienced academics, such as C.\ Sunstein and M.\ Hiscox advised on the unit's launch, with M.\ Hiscox later becoming the founding director of the unit. The main mission of BETA is "to improve the lives of Australians by generating and applying evidence from the behavioural and social sciences to find solutions to complex policy problems". By designing and testing various possible solutions, BETA tries to influence the way in which policies are shaped so that they better correspond to the needs of the population. BETA currently employs 27 staff members, including economists, psychologists, research methodologists (RCT specialists), data and statistics analysts, and policy experts. BETA also constantly collaborates with the Academic Advisory Panel, which helps the unit with advice on their research and evaluation techniques.

Scope of Work and Methodology

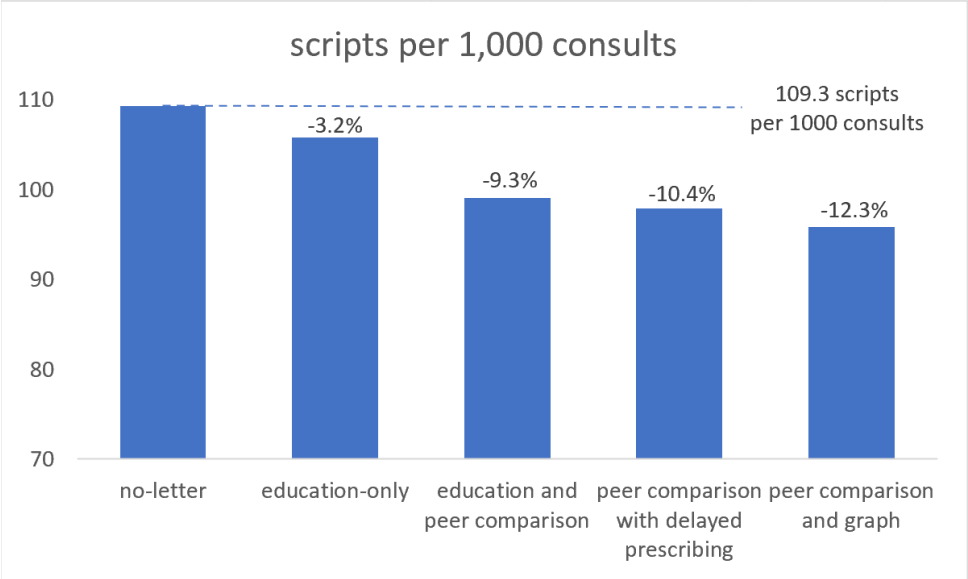
BETA is one of the nudge units that works in a range of policy domains, including economy, finance, environment, justice, safety, employment and many others. Since its creation in 2015, the unit has already finished many successful projects in all of these areas, proving that it can diversify its work without harming its effectiveness. While the BETA's methodology relies heavily on RCTs, the unit has developed a unique framework that helps to design policies that are later being tested. This framework, created by P.\ Ames and M.\ Hiscox, is called the "4D Framework", as it employs a four-phase methodology: discovery (identifying the existing policy problem), diagnosis (conducting pre-existing data to identify the appropriate interventions), design (designing the intervention and appropriate trials), and delivery (implementing and analysing the trial).

Randomised controlled trials play a crucial role in BETA's work. According to data from the Campbell Collaboration, Australia ranks fourth in the world in terms of the total number of RCTs in the field of social policy, surpassed only by the US, Canada, and the UK. The members of the Australian government are also very supportive of the wide use of RCTs in policy development. This explains why RCTs form the basis of BETA's methodology: the unit always rigorously tests the suggested policies before recommending them to the government. BETA also advocates for transparency and ethics in RCTs. Their approach includes publishing results regardless of outcome and adhering to national ethical guidelines and independent ethics reviews.

Example of a Completed Project

One of the most impactful BETA projects was reducing the overprescribing of antibiotics, completed in 2017–18. The project was of great significance, as antimicrobial resistance is now one of the greatest global health threats, causing many deaths annually worldwide.

The project is also a great example of how BETA uses RCTs to suggest appropriate policies. The four initially proposed solutions included sending out letters to GPs whose antibiotic prescription rates were among the highest in their regions. These letters used four different approaches: education, education with peer comparison, peer comparison with delayed prescribing (a strategy when a doctor tells the patient to only use antibiotics if the symptoms persist after several days and not immediately), and peer comparison with a visual bar chart of prescribing rates. "Peer comparison" in this case means stating in the letter that the GP to whom the letter is addressed prescribes more antibiotics than a certain percentage of all GPs in their region.



Main findings on antibiotic prescription levels following interventions

It is evident that peer comparison with a graph has had the most impact among the four solutions. The percentages shown on the chart are results after six months of the implementation. The figures in the follow-up report that covered the effect over a full year has shown that a total impact of approximately 190,000 fewer prescriptions across all four letter conditions was achieved.

This project can serve as an example of the cost-effectiveness achieved by BETA. These simple letters represented a minimal-cost intervention that had a substantial public health impact. The follow-up report also suggests that the effect is persistent and might potentially even strengthen over time. This makes this intervention highly effective.

Limitations

There were several projects completed by BETA that have raised many concerns and controversies. One such project, in which it participated in the development and refinement, was the Robodebt Scheme, an automated debt collection system. The problem appeared to be in the algorithm that used income averaging and thus didn't account for irregular work patterns, which

triggered false debt notices. The scheme was declared unlawful by the Federal Court in 2019, and it currently represents one of the most significant ethical and professional failures associated with using behavioural insights in Australia.

There are also other limitations that are applicable to overall performance of BETA rather than any particular projects. One of the main ones is associated with the use of RCTs as a primary testing method. The criticism includes concerns regarding external validity (RCT findings might be non-generalisable to real-world implementations at scale), potential Hawthorne effects (participants might change their behaviour knowing they are being observed), timeframe constraints (short-term results of RCTs might be very different from long-term effects once the policy is implemented), etc.

Peru

Overall History

The nudge unit of Peru, MineduLAB, was one of the first ones to be created in a resource-constrained context. As was mentioned previously, most of the existing research in the early 21st century was focused on designing interventions for WEIRD societies, which suggests that creating this behavioural insights unit was particularly challenging. Although currently the government of Peru is working on the development of several nudge units within various governmental departments, MineduLAB, created in 2013 within the Ministry of Education of Peru, was a pioneering initiative, often described as the first government innovation unit in Latin America to systematically use behavioural research and RCTs for policy innovation. Its success and approach later inspired the creation of other similar evidence-based structures in Peruvian government agencies.

Development of the Examined Unit

The reason for creating MineduLAB was the crisis in Peru's education system, when the country ranked last in the PISA (OECD Programme for International Student Assessment) ranking, which assesses the basic literacy skills of students from participating countries. This demonstrated the need for radical institutional innovation, and accordingly, MineduLAB was positioned within the Office of Strategic Planning and Monitoring. A defining characteristic of MineduLAB, the one that makes it different from other nudge units including those in Singapore, Australia and UK, is the severe budget constraints that the institution faces. In 2015–16, it operated with a team of only four employees, while having minimal funding for designing, testing and implementing interventions.

Scope of Work and Methodology

The main mission of MineduLAB lies in the scope of education; however, within it, the range of implemented policies is very broad. Apart from improving student learning outcomes, the unit also works on addressing problems such as teacher absenteeism, student dropout, and school maintenance problems.

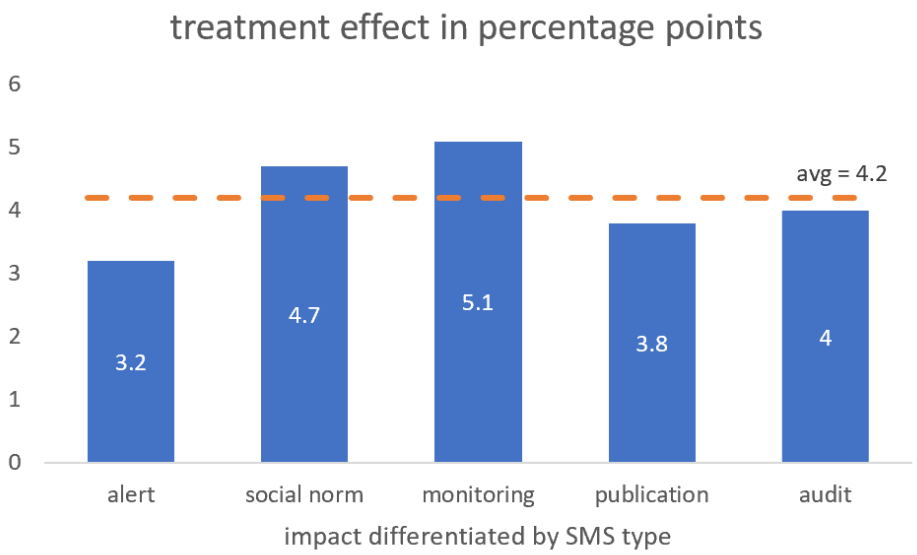
The financial constraints faced by MineduLAB largely determine its operating model, forcing it to rely on low-cost interventions and use existing data rather than conducting costly primary research. Nevertheless, the unit still relies on RCTs, as they are one of the most rigorous testing methods that can be used to evaluate policies, but strives to make them as inexpensive as possible for the unit, for example, by using low-cost intervention methods such as SMS-based interventions and utilising existing administrative data.

MineduLAB is also influenced by its partnerships with J-PAL (Abdul Latif Jameel Poverty Action Lab) and IPA (Innovations for Poverty Actions). The unit operates through a collaborative model involving government officials and international researchers, including those from J-PAL and IPA. This helps MineduLAB to gain access to valuable data and knowledge without incurring additional high costs. However, this structure was criticised for excluding Peruvian scientists and being influenced by certain external interests. This led to the creation of a more inclusive structure, in which the advisory board consisted of both local and international researchers.

Example of a Completed Project

One of MineduLAB's most successful interventions is "SMS PRONIED". Completed in 2015–16, this project targeted school maintenance programme compliance. The initial problem was in bureaucratic noncompliance, and the desired result was in nudging school administrators to complete the required procedure in order to get funds for infrastructure maintenance.

The project was completed using a large-scale RCT with several thousands of schools involved. MineduLAB tested several types of messages sent, and it was evident that all message types significantly improved compliance, with no approach proving superior. This is demonstrated in Figure 3 below.



Impact of various SMS types on the level of compliance by schools

The project was especially successful in terms of its cost-effectiveness, as for every $1 spent, approximately $800 of funds were reported by the school administrations in a timely manner. The programme continued to work, and by 2017 it was a national policy implemented for all schools in the country.

Limitations

One of the main challenges for MineduLAB is, not unpredictably, scaling. Despite being able to develop suitable policies in the experimentational process, the real-life implementation on a national, or at least regional, level was very difficult, with only 4 innovations out of 21 being successfully scaled. The main obstacle to implementing the proposed measures may lie outside MineduLAB's sphere of influence, as the barrier appears to be limited local capacity. Given that Peru is a developing country, the allocation of a limited budget is a serious challenge for the government, and not all projects are realised due to a lack of financial or human resources.

Another challenge seems to be the isolation from core political, policy and budgetary processes. In recent years, MineduLAB has significantly reduced its operations, suggesting that the unit is likely to some extent 'disconnected' from processes within the Ministry of Education and is unable to expand its activity due to limited autonomy and bureaucratic constraints.

Evaluation, Comparison and Conclusion

In all three examples, the BIT template has been adapted by the examined nudge units in different ways, and it is possible to analyse these differences in terms of several aspects. In terms of evaluation methods, Australia's BETA is closest to the 'traditional' model based on randomised controlled trials (RCTs), as they actively perform state-approved experiments in

social policy using this rigorous methodology; however, the unit also uses additional tools such as surveys, online experiments, and qualitative research when appropriate. Singapore's Innovation Lab is quite different in this sense, relying primarily on design thinking, ethnography, and rapid pilot projects rather than large-scale RCTs, still "experimenting", but in a broader sense. Peru's MineduLAB, in contrast, is RCT-focused, but in order to keep trials feasible under tight financial constraints, the unit uses RCTs in a low-cost, "lean experimental" form that primarily relies on inexpensive measures, such as SMS messages and the use of pre-existing administrative data.

The units studied also differ in their location within the government, which to some extent affects the degree of their interaction with or isolation from other government departments. The Innovation Lab is part of the central HR Public Service Division, which brings it close to interdepartmental capacity building but limits its formal authority over line ministries. BETA is a federal policy unit at the heart of government, with direct access to the cabinet agenda and broad authority across various sectors; and MineduLAB is located in the Ministry of Planning and Monitoring, close to education data and programme design, but somewhat isolated from major policy and budget decisions. This, in turn, determines resources and staffing: BETA's multidisciplinary team of about 27 specialists allows it to support large-scale trials and an experts' advisory board; the six-person Innovation Lab team prioritises facilitation and design support rather than data-intensive evaluation; and the four-person MineduLAB team must rely on partnerships with organisations such as J-PAL and IPA to compensate for severe resource constraints.

Pathways to scale and implicit definitions of "success" for the three units are also different. BETA's authority is based on its central positioning and on the credibility of rigorous RCTs, which makes it relatively easy for the unit to scale interventions to a national level successfully and to justify them politically in terms of measurable impact and cost-effectiveness. Innovation Lab, in contrast, discovered in its first years that acting as a central consulting unit produced "external" solutions with weak implementation, and has therefore shifted to a model in which agencies have their own project teams and the Lab provides coaching. This shapes the unit's definition of success, as for Innovation Lab, it is less about isolated impact metrics and more about embedding human-centred service design and cross-agency collaboration into routine practice. MineduLAB also uses a different model. Despite the success of some extremely cost-effective interventions, only a small fraction of its tested innovations have been successfully scaled, which is mainly due to limited state capacity, fiscal constraints, and organisational distance from core power centres. In this context, success will be defined in terms of improving bureaucratic compliance, strengthening administrative routines, and gradually building an evidence-based culture inside the ministry, as it is in terms of headline impact estimates.

Taken together, all these patterns suggest that the BIT template discussed at the beginning of the section (wide use of RCTs and central sponsorship) is applied only partially in the examined units. While some of its core "ingredients" (experimentation in some form, a connection with the government, and a narrative of behavioural "success") can be found in all three countries, the exact models of the units are still influenced by various exogenous factors, which lead to different

institutional locations, resource endowments and political logics, which also in turn affects the definition of an "effective behavioural governance".

V. Benefits of a National Nudge Unit

Abstract

Since the launch of the UK's Behavioural Insights Team (BIT) in 2010, behavioural science has become more embedded in policymaking, and many governments have established dedicated behavioural units. In this section, a "national nudge unit" is defined as a government-mandated behavioural insights unit with an ability to coordinate across different departments and agents within the government. It is a centralised unit that enables shared trial and evaluation protocols, common documentation and reporting standards, and cross-department governance that supports consistent and accountable application of behavioural tools (OECD, 2024).

This section evaluates how establishing such a unit can add value, and asks what the unit can enable that departments or ad hoc behavioural teams are not capable of delivering consistently. Three main claims of this section are as follows: (i) institutionally, a national nudge unit can strengthen governance agility, institutional memory, transparency, and ethical safeguarding through shared oversight; (ii) operationally, it can improve cost-effectiveness and implementation speed by reducing administrative friction while preserving choices for individuals; and (iii) culturally, it can improve the policy alignment by taking into account local behavioural diagnostics and contexts instead of transferring blanket policy from elsewhere. The subsequent part then discusses the spillover effects and distribution concerns brought by the unit. Finally, the section concludes with implications for policy delivery in increasingly data-rich contexts such as green and digital transformation.

Institutional Values

Institutional values represent the strategic and structural contributions of a national nudge unit, focusing on its role in enhancing the fundamental quality, integrity, and continuity of governance in the long term. A unit can function as intellectual and ethical infrastructure insofar as it puts concrete institutional mechanisms, including evaluation standards, oversight mechanisms, and documentation, into the policymaking process. The contributions of the national unit can be categorised into four dimensions: governance agility, institutional memory, transparency, and ethical safeguarding.

Agility and Rapid Response

A national nudge unit allows the state to complement longer policy cycles with a more proactive model of crisis management by maintaining rapid "test-and-learn" capacity. Zhu (2024) identifies that having a national nudge unit helps to facilitate rapid government interventions during times of crisis. The centralised unit can deploy "accessible, pre-vetted guides" that provide immediate strategies and coordinate messaging, in response to national emergencies such as energy surges and pandemics (Zhu, 2024). For example, behavioural insights in vaccination messaging where the public was informed to be the "top of the queue" were estimated to yield 42,000 additional first-dose vaccinations within the targeted cohort (BIT, 2025). This agility can ease bottlenecks when the public needs immediate remedy during an emergency. It also offsets the slower pace of departmental decision-making driven by clearance requirements, procurement constraints, and budgeting.

Behavioural Research and Development (R&D) Hub and Institutional Memory

A centralised unit functions effectively as a Behavioural R&D hub, providing a safe sandbox for the execution of Randomised Controlled Trials (RCTs) in which policies can be tested under limited-risk conditions—tested first through reversible, small-scale pilots in bounded settings with ethics review (Haynes, 2012). The unit's distinctive contribution is the trial infrastructure it maintains—standard protocols, analytical expertise, access to sampling and delivery channels, and practical know-how for embedding experiments in real services—which enables policymakers to "Test, Learn, and Adapt" interventions before national scale-up (Haynes, 2012). Through small-scale trials, the unit builds credible evidence which allows the policymakers to identify effectiveness and implementation feasibility of particular policies, as well as to detect null or adverse outcomes early. This R&D function contributes to increasing the evidential quality which underpins the implementation decisions, mitigating the risks of scaling ineffective or harmful policies.

Furthermore, as noted by Halpern and Sanders (2016), the units can capture essential "institutional memory". Under traditional government structures where there are no cross-department units like the national nudge unit, the lessons learned from successful interventions are only kept within the specific division of departments or ministries, or even lost after the disbandment of project teams. On the other hand, a national nudge unit can work as a platform where trial outputs are systematically recorded for future reference, such as trial registries, intervention designs, results including null and adverse effects, and heterogeneity findings. This allows the policymakers to build on success while avoiding the repetition of the errors seen previously. The inter-department nature of the national nudge unit might also allow further coordination between departments and ministries, promoting diffusion and reuse of proven intervention components.

Transparency and Standard Setting

Figure 1. MINDSPACE Framework Checklist

Messenger	we are heavily influenced by who communicates information
Incentives	our responses to incentives are shaped by predictable mental shortcuts such as strongly avoiding losses
Norms	we are strongly influenced by what others do
Defaults	we 'go with the flow' of pre-set options
Salience	our attention is drawn to what is novel and seems relevant to us
Priming	our acts are often influenced by sub-conscious cues
Affect	our emotional associations can powerfully shape our actions
Commitments	we seek to be consistent with our public promises, and reciprocate acts
Ego	we act in ways that make us feel better about ourselves

Source: Institute for Government

The MINDSPACE framework, mentioned in Dolan et al. (2010), serves as a checklist to systematically identify which behavioural levers are used in an intervention and where the policy requires evaluation, reporting and review. This framework states nine critical drivers of human behaviour: Messenger, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitments, and Ego (Zhu, 2024).

By mandating that every proposed intervention is checked against these nine criteria through required documentation templates and review checklists, a national nudge unit can standardise design choices and ensure that the rationale, delivery channel, and anticipated risks of interventions are recorded in a consistent manner (Zhu, 2024). This is important as some levers like "priming" or "affect" are often criticised for potentially having manipulative effects, and transparency of policies depends on making these mechanisms explicit. Moreover, clear and objective standards generate an audit trail that explains why certain behavioural policy was implemented, supporting transparency. The disclosure of the audit trail also allows academic critics and parliamentary oversight committees to evaluate and scrutinise the rationale, ethical acceptability and effectiveness of behavioural policies, fostering administrative accountability.

Ethical Safeguarding

A national nudge unit can provide an ethical governance function where it is mandated and resourced to do so. This helps to assess interventions that could shape an individual's choices that are often subject to concerns regarding manipulation. In practice, the unit can apply core principles, such as autonomy, welfare, transparency, and proportionality and consent, where feasible. It can then operationalise these through processes such as ethics review, disclosure and documentation standards, stakeholder consultations and safeguarding measures that enable opting out to be easy and straightforward. Although the contests around ethical acceptability will not be eliminated, these principles and measures can reduce normative concerns. As Thaler and Sunstein (2008) argued, the nudges are most defensible when freedom of choice is maintained. Having a national nudge unit can contribute to this through centralised ethical review, and potentially build public trust in the long-term that is fundamental for maintaining the policy effectiveness.

Operational Values

The operational values entail how the use of nudges can, in some settings, achieve efficiencies that traditional tools struggle to deliver without higher administrative burden. Traditional regulatory tools often involve high administrative burden with significant time and financial costs. However, by complementing such policies with behavioural insights, operational friction may be reduced; the nudge unit can design interventions using standardised experiments, translate findings to practical implementation, and help different departments adapt previously successful approaches to align with the specific policy purposes. This supports the improvement of quality and practicality of nudges used for public administration. Such operational values brought by the national nudge unit are categorised into fiscal cost-effectiveness, speed of implementation and preservation of individual autonomy.

Cost-Effectiveness and Budgetary Fit

Behavioural interventions, including the use of nudges, tend to cost far less than conventional government policies. As they only require minor tweaks to methods in which information is transferred to the public or choice architecture rather than significant infrastructure investment, they often "fit in the current budget allocation" (Zhu, 2024). This means that there is often less need for the government to find additional financial resources to fund behavioural interventions. Meta-analyses conducted by Mertens et al. (2022) demonstrated that choice architecture interventions achieve a "moderate effect size" comparable to more resource-intensive methods, such as direct financial incentives or large-scale education campaigns with significantly lower costs. Crucially, the cost-benefit ratio is higher for nudges due to their minimal implementation requirements, allowing the state to achieve meaningful gains under constrained budgets.

Furthermore, when complemented with other policies, nudges have potential to accelerate and amplify the effects of the original policies, making the overall return of investment even higher (OECD, 2017). For example, OECD (2017) reported that a UK RCT in which adding social-norm messages to HMRC's standard reminder letters for late self-assessment taxpayers increased the payment rates by 5.1% in the best sample, and brought forward around £9 million in revenue. This illustrates how a low-cost behavioural tweak can strengthen existing policy instruments

instead of substituting them. In fact, the national nudge unit can work as a bridge that enhances this complementary nature of nudges with policies in various fields.

High Yield in Short Periods

A national nudge unit can generate high policy yields within short timeframes. The unit specialises in interventions that remove administrative frictions in existing services; can be deployed through existing delivery channels without major procurement, budgeting, or statutory changes; have measurable impacts in administrative data; and can be quickly iterated via standardised trials. Particularly, the application of the EAST framework developed and disseminated by BIT operationalises this delivery logic by translating empirical evidence into policy delivery. For example, simplifying the court summons forms in New York City and making the whole process easier helped to reduce the failure-to-appear rates by 13% (BIT, 2024). BIT reports this reduction in failure-to-appear is associated with fewer arrest warrants being issued, suggesting that a small redesign to an existing form can produce substantial operational gains when scaled. Moreover, in an RCT conducted in three waves over 2.5 years, involving nearly 700 representatives from a Canadian government service agency, it was found that adding performance feedback as a nudge in organ-donation encouragement and reminder emails increased daily donation signups by 25% (House et al., 2024). This shows how a national unit can convert trial evidence into implementation-ready adjustments that can be replicated across similar administrative contexts, implying high yields under time constraints as well as the potential for institutional memory argued in Section 5.2.2.

Preservation of Autonomy

Unlike traditional mandates that rely on enforcement and the threat of punishment, nudges preserve freedom of choice for individuals. By incentivising citizens to make better decisions through nudges while ensuring that the path to opting out remains “easy and cheap” (Thaler & Sunstein, 2008), the state can avoid administrative resistance often caused by strictly enforced measures. This nature makes behavioural insights suitable for policy areas where direct enforcement is controversial and may be perceived as overreach, including public health and personal finance. This preservation of autonomy also allows the government to build public trust, contributing to long-term policy operations. However, it is important to acknowledge that the nudges are sometimes perceived as a form of covert influence. Hence, the national unit might act as a safeguard which ensures that its interventions are transparent and “easy to opt out”, minimising the risks of manipulation that the public might unconsciously be exposed to.

Cultural Salience

The effectiveness of behavioural science is not always universally consistent; it is dependent on the cultural habits and perceptions of the population being nudged. This sub-section addresses how and why effectiveness of nudges varies across cultures and the role that a national nudge unit plays in adapting nudges to local cultures and values. Particularly, the role of a national nudge unit in local testing, building culturally relevant behavioural diagnostics, and avoiding policy transfer errors will be discussed.

The WEIRD (Western, Educated, Industrialised, Rich and Democratic) Biases

Henrich et al. (2010) and Talhelm (2025) documented a profound bias in behavioural science research, stating that approximately 96% of samples are drawn from WEIRD countries. Given that only 12% of the world's population accounts for those living in Western countries, Henrich et al. (2010) and Talhelm (2025) warned that the science of interventions and policies are tenuous if they are built on that thin slice of humanity, implying that the published behavioural studies might not have wider applicability. Therefore, having a national nudge unit that conducts locally contextualised behavioural research helps mitigate a mismatch between external policies and the local need by generating context-specific evidence. This ensures that implemented nudges are not only scientifically sound but are culturally grounded in addressing specific issues of the local population.

Incentives and Norm-Based Levers across Different Cultures

The divergence in behavioural policy effectiveness across cultures is often visible in the relative performance of financial incentives against norm-based levers. Cross-national evidence suggests that the relative effectiveness of behavioural levers varies across cultural contexts. In Western countries like the US and the UK, market-based incentives involving financial gains tend to be a better motivator for behavioural changes while in non-Western countries like China or Mexico, social norms and reputational triggers tend to play greater roles in influencing public behaviour (Talhelm, 2025).

For example, in China, paying students money to study did not increase their scores in exams. In fact, "at one of the three schools in Shanghai, paying students seemed to decrease performance" (Gneezy et al., 2019; Talhelm, 2025). Conversely, paying students led to increased attempts to answer questions and better overall scores in the US, demonstrating that the marginal impact of monetary incentives can vary by context. Talhelm (2025) also found that money was a more effective motivator for behavioural changes in English than Hindi. Therefore, these comparisons do not imply that nudges are inherently more effective in non-Western contexts, but they indicate that the relative returns to behavioural levers such as incentives and norms vary depending on the culture, providing support for the local behavioural experiments and testing over direct policy transfer.

Universal Mechanisms and Cultural Differences

Certain psychological mechanisms, such as risk-aversion, appeared to be universal with little variance between nations. Using over 12,000 samples from Japan, Canada, and the US obtained through an online survey, it was shown that "risk-averse attitudes toward air pollution resulting from industrialisation were significantly moderated by the [behavioural] intervention" (Komatsu et al., 2022). However, the triggers required to activate such traits differ between countries. Zhu (2024) found that the Japanese population were highly reactive to the feeling of being observed by the community when making decisions. In Kyoto, a poster of "watchful eyes" at a Kyoto intersection appeared more effective than fine-based deterrence at stopping illegal parking. Zhu (2024) suggested that this outcome is likely caused by the trigger of the specific cultural fear of reputational loss, instead of financial loss. In Kenya, the effectiveness of cash transfers were

amplified when it was complemented with messaging emphasising community harmony rather than individual financial gains (Talhelm, 2025). These examples imply how framing that takes cultural tendencies into account can enhance the policy outcomes.

The Role of a National Nudge Unit under Cultural Differences

The national nudge unit performs strategic localisation by refining universal behavioural principles to align with local values and demands. Under the MINDSPACE framework, individualistic cultures may be more responsive to "Ego" or personal gains, while communal cultures are better served by emphasising "Norms" or collective responsibility. Talhelm (2025) argues that adapting interventions to culture improves the effectiveness of nudges, and by centralising such intervention through the national nudge unit, the cultural salience of choice architecture is more likely to be achieved. However, it is also important to acknowledge the existence of within-country heterogeneity and risks of stereotyping upon implementation, suggesting the need for empirical testing instead of oversimplifying the policies based on assumptions that might not align with the diverse populations. Thus, to further improve cultural salience of policies, the national unit may conduct localised RCTs across different sub-demographics and regions, building a choice architecture that resonates with the target population.

Indirect Effects of Implementing Nudges

Spillover Effects

To evaluate the effectiveness of behavioural interventions, it is critical to consider unintended spillover effects that can occur in complex cognitive environments. Koch et al. (2023) identified that reminder nudges can cause crowding out where increased attention to one policy or behaviour leads to the neglect of other behaviours, resulting in negative spillover effects overall. One hypothetical example would be that a nudge designed to increase recycling rates might reduce efforts in energy conservation as citizens might think that they have already fulfilled their moral obligation to act in an environmentally friendly manner. Reduced efforts in energy conservation might consequently diminish or even offset the positive effects brought by the nudge implementations.

Furthermore, the persistence in the effects of interventions is challenged when the choice architecture is withdrawn. Koch et al. (2023) found that while the positive effect on reminded actions diminishes after reminders are withdrawn, negative spillovers on nonreminded actions remain persistent. This suggests that reminder-based interventions may not reliably support habit transformation or strengthen intrinsic motivation; behavioural changes may be tied to the continued presence of the prompt rather than internalised routines (Talhelm, 2025). Although the challenges of offsetting spillover effects still remain, having a centralised nudge unit might be helpful to maintain the consistency of nudge policies and their integration.

Impact on Equity: Nudges' Regressive Nature

The use of nudges by a government body is also contested under the ethical matter of equity and equality. Ghesla et al. (2020) conducted a study on distribution effects of default nudges used for

electricity options in Switzerland. It was found that while the default nudge is successful at curbing greenhouse gas emissions, it leads poorer households to pay more for their electricity consumption than their willingness to pay while richer households pay significantly lower for green electricity than the fee that they are willing to pay (Ghesla et al., 2020). Although some might argue that all households have choices to opt out from such default settings in the electricity retail market, many fail to do so due to multiple cognitive biases, including inertia or complexity under limited time available for decision-making (Samuelson & Zeckhauser, 1988). In fact, the regressive effect is likely to depend on the distribution of preferences and the ability to opt out which might correlate with the public's education level, time limits and digital access. As a result, the poorer households remain in a situation where they face higher costs for electricity than their willingness to pay while the wealthier households enjoy benefits, suggesting that the real income inequality might widen as a result of nudge implementation (Sunstein, 2022).

Conclusion

In conclusion, a national nudge unit can offer strategic values that exceed the sum of the impacts of individual behavioural interventions. A centralised body that operates a variety of nudges across different departments of government is important to establish systemic governance. Institutionally, it helps to ensure agility, captures memory from past campaigns, and upholds ethical standards. Operationally, it proves a cost and time efficient alternative to conventional mandates through improved choice architecture and budgetary fit. In addition, the unit can act as a cultural bridge when domesticating behavioural policies that have worked elsewhere in the world. The WEIRD bias and contrasting incentive structures in Western and non-Western contexts showed that policies cannot simply be imported; the national unit can tailor policies to suit the culture and contexts of the designated country. Finally, a national unit is well placed to monitor the complex indirect effects that might occur as a result of nudge implementation and put in place preventative measures for negative spillovers or regressive distributional impacts. Hence, having a national nudge unit is not only important for securing positive policy outcomes, but also for future development of countries as governments pursue green and digital transformation, where behavioural tools can accelerate adoption and compliance.

VI. Limitations and Criticisms of a National Nudge Unit

The Ethics of Nudging

Introduction

Since the publication of *Nudge* by Thaler and Sunstein (2008), behavioural public policy has been accompanied by sustained ethical debate, particularly concerning autonomy, manipulation, and the appropriate limits of state influence. This section does not seek to resolve these disputes in the abstract. Instead, it adopts a policy-focused position: nudges can remain a legitimate and valuable tool for governments, but only within clearly defined ethical boundaries.

We argue that nudge units should prioritise interventions that support broadly shared and uncontroversial goals, reduce friction and error, improve access to information, and help individuals follow through on their own stated intentions. Nudges that simplify processes, correct distorted choice environments, or strengthen individuals' capacity for reflective choice are generally ethically defensible. By contrast, nudges become more problematic when they rely on opaque, reason-bypassing mechanisms in morally contested domains, implicitly select citizens' values rather than supporting their own, or impose unequal burdens on vulnerable groups.

Accordingly, this section treats nudging not as a single policy instrument but as a spectrum of interventions with varying ethical risks. It distinguishes defensible forms of means-paternalism from more controversial forms of value-paternalism, and develops practical guidance for how nudge units should design, evaluate, and limit their interventions within liberal democratic policymaking.

Nudging, Choice Architecture, and Libertarian Paternalism

The ethical debate begins with the definition of nudging itself. Thaler and Sunstein define a nudge as 'any aspect of choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives' (Thaler and Sunstein, 2009, p.697). Nudges are thus defined negatively, insofar as they fall short of coercion, prohibition, or material inducement, and must be 'easy and cheap to avoid'.

This definition separates nudges from coercion by framing them as 'freedom-retaining' (Siipi, p.697). Nudges are therefore libertarian insofar as individuals remain free to choose otherwise: no option is removed, and no financial penalty is imposed (Barton & Grüne-Yanoff, 2015, p.344; Leggett, 2014). Nudging is paternalistic only in the sense that it aims to make people 'better off, as judged by themselves' (Thaler and Sunstein, 2009, p.698).

The justification rests on behavioural economics: individuals are systematically boundedly rational and prone to inertia. Nudges, on this view, help people achieve outcomes they would endorse under conditions of full information and deliberation. They fulfil private welfare, defined as the satisfaction of an agent's true preferences, even if these preferences are unbeknown to them (Barton & Grüne-Yanoff, 2015, p.346; Sunstein and Thaler, 2003). Nudges thereby act as a form of means-paternalism, as they do not redefine the nudgee's ends, but rather the means by which those ends are attained.

However, the 'as judged by themselves' standard is also where a core ethical vulnerability emerges. In many real-world domains—particularly those that are morally contested, identity-linked, or culturally heterogeneous—policymakers cannot straightforwardly observe what counts as someone's "true" preference. Thaler and Sunstein justify paternalism as a means to 'influence choices in a way that will make choosers better off, as judged by themselves', had they possessed 'complete information, unlimited cognitive abilities, and complete self-control' (Thaler and Sunstein 2009, p.698). Yet precisely because this benchmark is counterfactual, it can be difficult to apply without contestation. In domains where preferences are stable and broadly shared (e.g.\ avoiding fines, completing forms correctly, reducing unintentional arrears), "as judged by themselves" is more plausible. In domains where preferences are plural and value-laden (e.g.\ organ donation, end-of-life choices, moralised pandemic behaviours), the benchmark becomes ethically fragile. This distinction helps explain why the same nudge mechanism may be ethically benign in one context and ethically fraught in another.

According to this definition, nudges rely on an outcome-oriented conception of rationality, 'irrespective of whether the process by which this choice was produced is rational' (Barton & Grüne-Yanoff, 2015, p.345). This tension between outcomes and processes motivates much of the autonomy-based critique. Critics generally accept that choice architecture is unavoidable, as choices must be presented in some form. However, the ethical significance of intentional behavioural steering remains contested. As Barton and Grüne-Yanoff (2015) note, nudging has evolved from a descriptive insight about decision environments into an explicit policy tool, and it is this intentional use of behavioural influence that raises normative questions. Nudge units should be most confident using nudges in domains where ends are widely shared and easy to specify, and more cautious where the state would in effect be choosing contested ends.

Autonomy and Non-Rational Influences

The most prominent ethical objection to nudging concerns autonomy. For critics such as Hausman and Welch (2010), nudges threaten autonomy not by restricting choice sets, but by shaping decisions through non-rational mechanisms. This concern applies especially to so-called Type 1 nudges, which rely on shaping choice architecture to produce particular outcomes through non-cognitive processes, notably heuristics.

To make this debate operational for policy, it is helpful to distinguish three autonomy concepts that are implicitly in play across the literature:

Concept	Description
Thin autonomy / freedom of choice	Autonomy as having options and not being forced. A policy violates autonomy if it removes options, imposes penalties, or coerces. Many nudges look acceptable on

	this standard because choice sets remain and opting out is allowed.
Rational agency (procedural) autonomy	Autonomy as making decisions through deliberation and reflection. A nudge may violate this by bypassing individuals' reasoning processes, even while leaving the choice set intact.
Volitional autonomy	Autonomy as acting in accordance with one's endorsed second-order preferences. Some nudges can *enhance* volitional autonomy by counteracting private manipulation or correcting distorted choice environments.

These distinctions matter because debates about whether nudges respect autonomy often involve critics and defenders talking past each other. A nudge can be "freedom-retaining" in the thin sense while still undermining procedural autonomy.

Critics argue that there is a morally important distinction between rational persuasion, which engages individuals' deliberative capacities, and choice shaping, which exploits cognitive biases to steer behaviour. On this view, the concern is not that nudges are coercive, but that they undermine individuals' control over their own decision-making processes. When defaults, framing, or salience effects influence behaviour without engaging reflection, actions may reflect the priorities of the choice architect rather than the considered judgement of the chooser (Hausman and Welch, 2010). In this sense, nudges may infringe upon individuals' autonomy even when freedom of choice is formally preserved.

This concern is developed further by Schmidt and Engelen (2020), who distinguish between influence that engages rational agency and influence that bypasses it. Many nudges operate primarily through automatic cognitive processes rather than reflective deliberation. From this perspective, respect for autonomy requires not only preserving options, but also engaging individuals as reasoning agents. However, for deliberation to be possible, nudges must be sufficiently transparent to be recognised and avoided if desired (Barton & Grüne-Yanoff, 2015, p.347). Although Thaler and Sunstein emphasise transparency in principle, the application of nudge theory has at times been criticised for lacking it in practice.

Organ Donation

The ethics of organ donor registration provides a clear illustration of these tensions. MacKay and Robinson (2016) analyse opt-in, opt-out, and mandated active choice (MAC) systems through the lens of autonomy. Drawing on Thaler and Sunstein's account of status quo bias, they argue

that both opt-in and opt-out systems rely on reason-bypassing influence by exploiting defaults. Such nudges also rely on conformity bias, or the tendency to follow others' behaviour unquestioningly (Siipi, p.709).

Opt-out systems, adopted in the UK in 2020, automatically register individuals as organ donors while retaining freedom of choice by allowing them to opt out, typically through written or online procedures (MacKay and Robinson, 2016). This format has been shown to significantly increase donation rates due to status quo bias, whereby individuals are predisposed to 'stick with the current state of affairs or choose default options' (MacKay, p.3). However, MacKay and Robinson contend that effectiveness alone is not decisive. Decisions concerning post-mortem organ use involve a moral right, and exploiting cognitive biases in this context risks bypassing individuals' rational agency (MacKay, p.4). Preferences regarding organ donation are heterogeneous and often tied to religious or cultural beliefs, making appeals to hypothetical 'ideal preferences' particularly problematic.

Their preferred alternative, mandated active choice (MAC), is more coercive in a narrow sense—for example, in New Zealand, renewing a driver's licence is conditional on stating a donation preference—but autonomy-promoting in a deeper one, as it requires individuals to engage reflectively with the decision. This critique casts a different light on the UK experience. Prior to the introduction of opt-out legislation, the Behavioural Insights Team conducted large-scale trials testing emotionally framed messages on government websites. As Wright (2013) reports, small wording changes led to an estimated 100,000 additional registrations per year. While widely cited as a behavioural success, this case exemplifies the ethical trade-off identified by MacKay and Robinson: increased compliance may come at the expense of informed, reflective consent.

What MAC and opt-out systems share is difficulty in achieving uniform effects across heterogeneous populations. While opt-out systems formally preserve the option to withdraw, the friction involved in doing so is not experienced equally. Similarly, MAC systems may disadvantage individuals without driving licences. Although options are not removed, additional friction is introduced, and this burden is unevenly distributed. As Blumenthal-Barby and Burroughs argue, 'injustices' are often experienced by 'marginal persons through defaults' (p.4). Where opposition must be expressed in writing, less-educated or illiterate individuals face disproportionate obstacles (Blumenthal-Barby & Burroughs, p.4). Language barriers create similar vulnerabilities (Blumenthal-Barby & Burroughs, p.4).

Defenders of nudging respond that this critique relies on an idealised conception of autonomy. In practice, choice environments are already structured, often by private actors whose incentives are poorly aligned with individual welfare. Leaving individuals subject to such environments does not obviously better respect autonomy (Schmidt and Engelen, 2020). Public nudges may therefore correct rather than introduce autonomy-undermining influences.

Moreover, not all nudges operate at a purely non-deliberative level. Hansen and Jespersen's distinction between Type 1 nudges (primarily automatic) and Type 2 nudges (which engage reflection) complicates the autonomy critique. Interventions such as traffic-light nutrition labels or energy-use comparisons may support, rather than bypass, deliberation.

For example, Barton's (2013) analysis of tobacco health warnings, including graphic images depicting the effects of smoking, demonstrates how nudges can, under certain conditions, enhance autonomy. Barton distinguishes between the informational and persuasive roles of health warnings. Graphic warnings convey risk information in a socially equitable way, reducing informational disparities, while emotionally salient images address smokers' second-order desires to quit. Although such warnings appeal to non-deliberative faculties and therefore infringe autonomy in one sense, Barton argues that this infringement is offset by gains in self-rule. According to the World Health Organisation, health warnings on cigarette packaging are among the six key measures required to reduce smoking prevalence, given that smoking remains the leading cause of preventable death in the developed world (Barton, 2013, p.208). Such warnings also counteract the implicit advertising embedded in cigarette packaging, itself a form of private manipulation.

If autonomy is understood as 'the capacity to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distortive external forces', then such nudging may restore a more autonomy-supportive choice environment rather than distort it (Christman, 2011, in Barton, 2013, p.209). Similarly, environments are often 'shaped by private companies that use choice architecture ill-matched to our decision-making procedures' (Schmidt and Engelen). This is particularly evident in the food sector, where firms 'systematically trigger and exploit our psychological dispositions so that we overeat and overspend' (Schmidt and Engelen).

In such contexts, nudges may correct distorted environments and strengthen rational agency rather than diminish it. Health campaigns that require calorie labelling in large UK food outlets, introduced in 2022, arguably enhance volitional autonomy by equipping individuals with information necessary for deliberate decision-making. These Type 2 nudges support slower, more reflective choice rather than simply steering behaviour automatically (Schmidt and Engelen).

If a nudge unit is to remain legitimate, it should prioritise interventions that (i) are compatible with thin autonomy and (ii) do not systematically undermine procedural autonomy, and where possible (iii) strengthen volitional autonomy by supporting endorsed preferences. This is one reason to treat morally contested domains as higher-risk, especially when "as judged by themselves" cannot be reliably operationalised.

The extent to which nudges interfere with autonomy therefore depends on the specific intervention and context. It would be misleading to dismiss all nudges as infringements on autonomy, just as it would be naïve to endorse them uncritically. Ethical evaluation must proceed on a case-by-case basis, using autonomy criteria that are explicit rather than assumed.

Manipulation, Transparency, and 'Slippery Slopes'

Closely related to the autonomy objection is the charge that nudges are manipulative. Siipi (2025) argues that commonly used definitions of nudging and manipulation overlap to such an extent

that a clear conceptual boundary is difficult to sustain. Both involve intentional influence, operate without coercion, and are typically low-pressure and avoidable. If manipulation is understood as influencing behaviour without the target's awareness, then opaque nudges risk falling within it.

Siipi's analysis highlights institutional rather than purely individual risks. Once reason-bypassing influence is accepted as ethically permissible, policymakers may gradually expand its use, reclassifying increasingly intrusive interventions as benign nudges. Appeals to individuals' 'true preferences' under ideal conditions are particularly vulnerable to abuse, as policymakers may project their own values under the guise of correcting cognitive error (Siipi, 2025).

Given the centrality of this critique, it is useful to state a policy-operational definition: a nudge becomes manipulative when (i) it is designed to work by exploiting non-transparent psychological mechanisms and (ii) the affected individuals could not reasonably understand or contest the influence exerted, even if they were generally aware that "nudging" exists. This formulation allows policymakers to distinguish between (a) openly informational or deliberation-supporting nudges and (b) interventions whose effectiveness depends on opacity.

Transparency and democratic control are frequently proposed as safeguards (Schmidt, 2017, p.413). Thaler and Sunstein (2008) invoke a Rawlsian principle of publicity, arguing that nudges should be publicly defensible to citizens (MacKay, p.352). Critics such as Hausman and Welch (2010), however, contend that this is insufficient. Respect for autonomy requires that citizens be informed when the state is deliberately exploiting cognitive vulnerabilities, even if such disclosure reduces effectiveness. Moreover, even when individuals recognise that an intervention is being used, they may not understand the mechanisms by which it influences behaviour: 'to the extent that people are unable to fathom the underlying mechanisms that bring about behavioural change, transparency is reduced' (Hertwig & Grüne-Yanoff, p.981).

Concerns about manipulation and legitimacy become particularly acute during emergencies, when fear and uncertainty heighten vulnerability (Hausman and Welch, 2010, p.135). Ruda's (2022) reflection on behavioural policy during the COVID-19 pandemic illustrates this risk and highlights the need for 'the accountability of legislative and parliamentary scrutiny'. While broadly supportive of behavioural science, Ruda warns that fear-based messaging may have long-term consequences, including erosion of institutional trust, that are not captured by standard RCT-based evaluations.

As the tobacco example demonstrates, emotionally salient messaging is not necessarily illegitimate: Barton (2013) explicitly defends warnings that are vivid and affect-laden, partly because they counteract private manipulation and support second-order preferences. The ethical risk during crises is therefore not "emotion" per se, but whether emotional leverage becomes a substitute for public justification and accountable policymaking, and whether behavioural techniques are deployed in ways that are hard to scrutinise or reverse.

Pandemic messaging informed by behavioural insights often appealed to fear or moral pressure. In collaboration with several American cities, the Behavioural Insights Team supported initiatives evaluating the most impactful COVID-19 messaging slogans, many of which relied on appeals to duty and fear. Ruda's critique reinforces slippery-slope concerns: once subtle state

influence becomes normalised, crises may licence far more intrusive behavioural control without adequate democratic scrutiny. Nudge units should treat emergency contexts as a distinct category requiring heightened safeguards: clearer lines of responsibility, explicit public justification, and evaluation metrics that extend beyond immediate behavioural proxies to include effects on trust and legitimacy (Ruda, 2022).

An Ethical Contrast between Nudges and Boosts

One influential response to ethical concerns surrounding nudging is the proposal to complement nudges with boosts. Hertwig and Grüne-Yanoff (2017) distinguish nudges, which steer behaviour by modifying choice architecture, from boosts, which aim to strengthen individuals' decision-making competences. Whereas nudges exploit cognitive biases such as inertia or loss aversion, boosts cultivate skills such as statistical literacy or simple decision heuristics.

From an ethical perspective, boosts address several autonomy-based objections raised against nudging. Critics such as Hausman and Welch (2010) object that nudges bypass deliberation; boosts, by contrast, explicitly engage reflective capacities and more closely resemble rational persuasion, with potential long-term benefits. This responds to concerns that nudges may undermine individuals' moral independence (Schmidt and Engelen, p.5). As Hausman and Welch note, 'unlike constraining someone or substituting your judgement for theirs, providing information and giving advice treats individuals as fully competent decision-makers' (p.127).

Such interventions increase the neutrality of choice architecture through education and training (Barton & Grüne-Yanoff, p.347). While choice architecture is inevitable, boosts—such as improving statistical health literacy or financial reasoning—aim to shift individuals from automatic to reflective decision-making (Hertwig & Grüne-Yanoff, p.977; Ruda, 2022).

As nudges must, by definition, be easily avoidable, their effects are reversible. Boosts, by contrast, have an educative function and must be transparent to the individual (Hertwig & Grüne-Yanoff, p.977). They are therefore necessarily transparent in both aims and mechanisms, reducing concerns about covert manipulation. Sunstein has argued that nudges and boosts are not mutually exclusive. In 2016, he introduced the notion of educative nudges, arguing that 'some of the best nudges are boosts' (Sunstein, 2016, p.10), precisely because they engage Type 2 reasoning rather than relying solely on heuristics.

Unlike Thaler and Sunstein's framework, however, boosts place explicit emphasis on empowerment (Hertwig & Grüne-Yanoff, p.981). As Leggett warns, a 'nudging state' risks becoming 'another voice trying to grab consumer attention in an already crowded market', indistinguishable from private sector marketers. In this context, boosts, by reasserting the importance of an active and educative state, may offer more empowering models of citizen engagement with longer-term benefits.

Nevertheless, boosts are not a universal substitute. They often require sustained engagement and cognitive effort, and may be less effective in time-pressured or low-salience contexts where

nudges perform well (Schmidt and Engelen, 2020). The ethical choice between nudges and boosts is therefore context-dependent rather than absolute.

Conclusion

The ethical debate over nudging does not yield a single decisive objection, but rather a complex set of trade-offs. Critics converge on concerns about autonomy, manipulation, and legitimacy, yet diverge in their normative foundations and practical conclusions. Defenders respond by emphasising the inevitability of choice architecture, the diversity of nudges, and the potential for autonomy-enhancing interventions.

For policymakers, the central takeaway is not that nudge units should be abandoned, but that their remit should be ethically disciplined. Nudges are most defensible when they operate as means-paternalism (helping people achieve widely shared ends) and when they are compatible with robust autonomy standards: preserving freedom of choice, avoiding systematic reason-bypassing in contested domains, and, where possible, supporting reflective decision-making. Conversely, nudges become harder to justify where "as judged by themselves" cannot be credibly operationalised, where mechanisms depend on opacity, or where interventions impose unequal burdens across vulnerable groups.

Accordingly, nudge units should prioritise transparent, deliberation-supporting interventions (including educative nudges and boosts) and apply heightened safeguards in ethically sensitive or crisis contexts. Greater attention to boosts, deliberative engagement, and institutional trust can help reconcile behavioural effectiveness with respect for autonomy and democratic legitimacy.

VII. Policy Implications: How Should Nudge Units Be Structured?

Introduction

When assessing whether nudge units still work, it is important to analyse their institutional characteristics. Institutional design can be integral to a unit's success, as it shapes the unit's reach, impact and longevity. This is because design choices determine how a unit is structured in terms of staff and stakeholders as well as how it operates in practice, including how it attracts work and implements effective policy interventions.

This section first examines the history of the Behavioural Insights Team (BIT) in order to understand its successful formation and development. Next, a framework to outline how to successfully design a nudge unit will be constructed and compared with the institutional design of BIT. Finally, this section highlights the importance of context, particularly institutional and political, and the unit's motivation because the framework derived is only meant as a useful starting point and should be adapted to each context. For example, directly replicating BIT's design and development would be ineffective. Instead, lessons from BIT's experience can be considered and adapted as necessary when forming a new nudge unit.

Formation and Development of the Behavioural Insights Team (BIT)

The first official nudge unit was the Behavioural Insights Team (BIT), which was set up by the UK Cabinet Office in 2010. Since December 2021, BIT has been owned by Nesta, an innovation charity. According to BIT's website, as of January 2026, they have 220 staff, 7 offices worldwide and have delivered 1800 projects world-wide. For the purposes of this section, success is not evaluated in terms of the effectiveness or societal impact of specific nudges, as this has been addressed earlier. Instead, success is defined using three metrics: policy value-added, cost-effectiveness and institutional resilience. Here, policy value-added captures the scope, uptake and influence of the unit's behavioural policies, while the other metrics assess whether the unit is economically viable (cost effective) and sustainable over the long term (institutionally resilient). Using these metrics, BIT's continued expansion and development demonstrate that it has been successful, providing a compelling case for examining the institutional design features that enabled this growth and adaptation.

BIT has gone through three distinct phases of development. Using the terminology of Neuhaus and Curley (2022), these phases are Proof-of-Concept (2010–2012), Global Roll-Out (2013–2021) and Monetarisation (2021–present).

However, before delving deeper into these phases, it is important to understand how BIT was established in the first place. As highlighted by Neuhaus and Curley (2022), in 2002, the UK Prime Minister, Tony Blair, set up the Cabinet Office Strategy Unit (COSU). The mission of this unit, as described by Blair, was to "Look ahead at the way policy would develop, the fresh challenges and new ideas to meet them" (Blair, 2010, p.339). The work mainly focused on problems and potential solutions generated by psychological insights. The application of behavioural insights within the government was further driven by the publication of *MINDSPACE: Influencing Behaviour for Public Policy* (Dolan et al., 2010). This was an influential publication and helped make the case for BIT's establishment, as well as providing an important framework used by BIT to apply behavioural science to the policy making process. Shortly after the publication of the MINDSPACE report, the new Coalition government, with David Cameron as the Prime Minister, set up BIT to satisfy the pledge to find "intelligent ways to encourage, support and enable people to make better choices for themselves" (Behavioural Insights Team, 2011). This was the beginning of the first phase.

Proof-of-Concept Phase (2010–2012)

BIT was established in 2010 with the primary objectives, which would then be assessed in the July 2012 sunset review (Behavioural Insights Team, 2011):

- Transform two major areas of policy, plus support work in a number of other policy areas as agreed with the Steering Board
- Spread understanding across government, including the use of behavioural approaches as an alternative or complement to regulation or bans
- Achieve at least a 10-fold return on the cost of the team

By the time the sunset review came around, BIT had "led six major pieces of work, published four policy reports, and contributed to tens of other policy areas; helped greatly to improve Whitehall's understanding of behavioural insights through joint projects, conferences, seminars, workshops and support in developing randomised controlled trials, and achieved savings of around 22 times the cost of the team and identified specific interventions which will save at least £300m over the next 5 years" (Behavioural Insights Team, 2012). Therefore, BIT passed its review and was allowed to continue to operate.

This sunset review provides transparent results from BIT's proof-of-concept phase, which can be evaluated against the success metrics introduced earlier:

Policy value-added: BIT's work influenced how policy was designed and implemented across multiple government departments, while also establishing methods and practices that departments could adopt. Its efforts to teach Whitehall teams how to apply behavioural insights expanded the reach and uptake of its policies, reinforcing its influence within the government.

Cost effectiveness: BIT delivered savings of roughly 22 times its operational cost, demonstrating the efficiency of its interventions. The projected savings over the following five years further underscore the lasting economic benefits of its policies.

Institutional resilience: BIT's establishment under the Coalition government built on what Labour had instituted through the COSU, thus highlighting its political resilience. Passing the sunset review reinforced this stability and emphasised the sustained support from senior stakeholders.

Alongside these sunset review objectives, Neuhaus and Curley (2022) highlight four more aims of BIT during this phase: demonstrating effectiveness, becoming the "epistemological authority" in behavioural public policy, promoting behavioural approaches through publications and networking, and setting methodological standards (e.g.\ via the 2012 paper "Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials").

These aims clearly highlight BIT's policy value-added on a global scale, stemming from its unique position as the first nudge unit and its obligation to showcase the potential of behavioural insights. This helped BIT to establish itself as the "global leader, if one wants to learn how to use nudges" (Oliver, 2013). BIT also created "the market for behavioural public policy ... [and] created

demand within that market" (Neuhaus and Curley, 2022). These factors helped lead BIT to its next phase of development.

Phase 2: Global Roll-Out (2013–2021)

BIT was part-privatised in 2014 after the sunset review concluded that BIT "should be given greater freedoms and flexibilities to respond to the growing demand for the application of behavioural insights, both within and outside government" (Behavioural Insights Team, 2012). BIT therefore became a social purpose company, allowing it to enter the free market economy and it was agreed that the cabinet, BIT's staff and NESTA would each hold a third of the company's equity (Neuhaus and Curley, 2022).

Press releases from the time help shine a light on the reasons and benefits of BIT's spin-out. David Halpern, the head of the unit at the time, stated that the "joint-venture framework" gave BIT the flexibility "to work for public bodies and foreign governments" without using UK taxpayer money to fund these ventures (Wintour, 2014). Wintour (2014) also cited Geoff Mulgan, former CEO of Nesta, who said that the partnership allowed for talent sharing as well as international expansion. Similar sentiments were shared by Service (2014), who said that the spin-out gave them the opportunity to take on work commissioned by UK government departments, foreign governments and private sector companies, when there is an underlying social purpose.

From the facts, as outlined by Neuhaus and Curley (2022), it is clear that BIT benefitted from this spin-out. BIT expanded the number of branches it had to places where their expertise were demanded on a long-term basis. These branches are located in North America, Oceania, Asia and also the Latin America and Caribbean region. BIT's revenue increased, with around 40% of its £14 million annual revenue in 2017 being accumulated outside of the UK (Quinn, 2018). BIT's staff also expanded a lot, increasing from just seven employees to a couple hundred by late 2019.

Once again it is important to evaluate BIT's development against the success metrics:

Policy value-added: the spin-out allowed BIT to work with a broader range of stakeholders, increasing the scope and uptake of its behavioural insights globally.

Cost effectiveness: greater autonomy over spending allowed BIT to operate efficiently, with public funding focused on projects that directly benefited the UK government while commercial work satisfied external demand.

Institutional resilience: spinning out gave BIT part-independence from the government, reducing its exposure to political volatility, strengthening its stability and ensuring long-term sustainability.

Therefore, the spin-out seems to have been a necessary response to demand for behavioural insights and helped BIT successfully grow. This led BIT to what Neuhaus and Curley (2022) describe as its Monetarisation phase.

Phase 3: Monetarisation (2021–present)

Nesta acquired BIT for £15.4 million in December 2021 (Nesta, 2021). This has allowed Nesta and BIT to become more intertwined. The reasons for this were summarised by David Halpern as he highlighted that the acquisition allows behavioural science to be blended with "innovative mixed methods, including data science, design, social psychology and collective intelligence". This will help "both organisations to better deliver what we were created to do—to harness innovation and evidence to make the world a better place" (Nesta, 2021).

This phase grants BIT greater autonomy from the government while also enhancing skills sharing and operational synergies through its full integration with Nesta, allowing for continued innovation. Although the term monetarisation may imply a shift towards commercialisation, BIT has remained a commercial social purpose company, with any profits reinvested into social impact initiatives (Nesta, 2021). While a common concern with private organisations is their tendency to take a short-term view (Caldwell, 2018), gaining full autonomy from the Cabinet Office may instead benefit BIT by reducing its exposure to political constraints, thereby allowing it to plan more effectively for the longer-term. Moreover, as a commercial social purpose company, BIT continues to face a profit motive, which can facilitate innovation and support its growth, helping to ensure that the organisation is positioned sustainably for the long term.

Evaluated against the success metrics, this phase represents a consolidation rather than a radical development like the previous phases. Full autonomy allows BIT to continue expanding the scope and influence of its work, reinforcing its policy value-added, while its social purpose model and financial independence support both cost-effectiveness and institutional resilience in the long term.

Overall, BIT has repeatedly adapted its organisational design in response to growing demand, enabling it to sustain its activities and extend its expertise globally. Across its three phases of development, BIT's core commitment to applying behavioural insights for social good has remained intact, despite significant institutional change. Thus, evaluated against the success metrics of policy value-added, cost-effectiveness and institutional resilience, BIT demonstrates how a nudge unit can be designed and evolve successfully over time. This makes it a valuable case for examining its institutional design choices.

Necessary Components of a Nudge Unit

The success of the world's first nudge unit demonstrates that careful institutional design is crucial for a nudge unit's effectiveness. A practical framework for achieving this has been identified by Halpern and Sanders (2017), which drew from their experience and knowledge, particularly of BIT. They summarised this framework with the mnemonic: APPLES (administrative support, political support, people, location, experimentation, and scholarship) providing a clear basis for institutional design considerations. This framework can be enhanced by including propositions taken from the experience of the Behavioural Insights Group Rotterdam (BIG'R) (Dewies et al., 2023).

Administrative support outlines the need to have senior level buy-in within the system. This can help give the unit more credibility and can help ensure the unit has the impact and support it deserves. For example, the cabinet secretary, the most senior civil servant of the UK government, supported BIT, which helped signal its legitimacy and potential to the rest of the government.

Political support is also a necessary component when setting up a nudge unit. Ensuring the unit and its aims fit with the current political environment and concerns of the government is integral for a nudge unit to be successful. This was crucial for BIT's formation as Prime Minister David Cameron and the Deputy Prime Minister Nick Clegg had an interest in behavioural insights so were willing to support BIT's creation.

People relates to the need to form a team with the right mix of skills and expertise. Each member of the team should possess at least one of the following key skill sets: an "understanding of government, knowledge of behavioural science, knowledge of policy and intervention design, analytical skills, interpersonal communication skills, and management skills" (Halpern and Sanders, 2017). Having the right team not only helps ensure a nudge unit can produce effective policy recommendations supported by robust evidence but also enables the unit to build strong relationships with key allies, helping ensure the unit's continued support and development.

Location is also important to ensure the nudge unit is close to the action. The nudge unit should be close to the people and institutions it wants to work with. Being close to the action can ensure the unit's work is integrated within the places it aims to help and so can reduce the risk of the unit becoming overlooked. It also can help the unit produce more effective work as they will have a greater understanding of the inner workings and inefficiencies of the institutions it works with.

Experimentation addresses the need to use empirical methods to demonstrate the unit's value and show that the policies work by quantifying its impact. For example, BIT uses the framework "test, learn, adapt" when approaching experimentation.

Scholarship means that it is important to know the behavioural literature and understand the challenges that you will encounter. It is also important to form an advisory group of academic experts, including psychology experts, to ensure that the unit is always up to date on new practices and ways of thinking. This helps ensure the unit's longevity and enables it to remain an effective asset for the institutions that rely on it.

A similar sentiment to this framework was echoed by Dewies et al. (2023), which looked at the Behavioural Insights Group Rotterdam (BIG'R) and derived a set of propositions to "describe and improve the integration of behavioural insights into public policy and administration". Five of these propositions are important to the design of nudge units.

Firstly, "competencies for conducting or managing research and knowledge about behavioural science are key for policy professionals working within BITs". This echoes the component Scholarship in the APPLES framework and is a lesson learnt from BIG'R, who didn't take into account the resources required when hiring advisors, leading to a high turnover of advisors.

Secondly, "team stability and/or good handover to new team members are key to the completion of policy cases, and to improve group learning and development". High turnover can slow down the workings of the unit, as policy cases tend to take place over a long period of time.

Thirdly, "effective boundary workers are important to increase reach and complete policy cases". This emphasises the need to bridge the gap between the nudge units and the institutions it aims to support. This links to People and the need for team members with a range of skills including interpersonal and management skills.

Fourthly, "BITs need to adapt their approach to policy cases (e.g., research methods, own role) to operate under different administrative circumstances and with different target behaviours". This indicates that the policy process must be sufficiently flexible to adapt to variation in policy cases and institutional contexts.

Finally, "to increase their achieved scope of change, BITs require a broad mandate that includes support for implementation". Having a broad mandate allows nudge units to "think big" and ensures that they have the resources they require to implement their policies.

Successful Institutional Design of BIT

The institutional design of the UK BIT is a key factor in its success, enabling it to act as a change agent and innovate effectively internationally as well as within the UK government. John (2014) presents this idea through the lens of organisational theory, examining the conditions under which organisations become more innovative. John (2014) focuses specifically on the concept of skunkworks, which are small units that pioneer innovations, which subsequently diffuse throughout the wider organisation. He identifies several conditions that facilitate the success of skunkworks, including:

1. Less hierarchy
2. Operate within a different framework of performance evaluation, where the unit is not subject to short-term management objectives
3. Are nurtured by senior manager champions who can protect them from turf wars
4. Are funded differently and have a separate structure of cost control
5. Occupy a separate physical space
6. Are subject to a longer time cycle for the measurement of success
7. Have low staff turnover so as not to disrupt the flow of ideas and memory of the organisation. A small group is important.

John (2014) highlights BIT's non-hierarchical structure, lack of a strict framework governing their performance and scope, senior-level buy-in from the Prime Minister and low staff turnover. Thus BIT satisfies four of the seven skunkworks conditions. The other three conditions are less directly applicable: the time cycle is not applicable as BIT was initially set up with a two-year sunset clause; BIT's non-separate physical space has arguably helped its success by enabling closer integration with government; and in terms of funding, since becoming private the financial

flexibility of BIT has likely improved as shown by the increased global reach of BIT branches and projects.

Overall, BIT's institutional design aligns closely with the characteristics of a skunkworks, supporting its role as a change agent both within the UK government and internationally. These characteristics also align closely with the extended APPLES framework, reinforcing its practical relevance for understanding successful nudge unit design.

Importance of Context

It is clear that the institutional design of BIT has been instrumental to its success and that this design can be generally summarised by the APPLES framework, with the BIG'R propositions enhancing the design considerations. However, it is important to recognise that each nudge unit is formed and designed within a unique context, particularly in terms of a country's institutional and political context.

The OECD (2017) analysed a collection of global nudge units and identified three institutional models: diffuse, central steering and project models.

The **diffuse model** is where existing units in a department or specialised agency in either the central or local government apply behavioural insights.

The **central steering model** is where a specialised unit, usually at the centre of government, focuses on "applying, supporting and advocating" (OECD, 2017) the use of behavioural insights across the government.

The **project model** is where specialised teams use behavioural insights for specific initiatives.

The OECD highlights that these models aren't mutually exclusive, they can co-exist and alter over time. BIT has developed from a central steering model to more of a diffuse model with external central steering support. Institutional make-up, administrative culture and the desired degree of use of behavioural insights are key determinants of which model is most appropriate.

Conclusion

As stressed throughout this section, institutional and political context, as well as a unit's underlying motivations, play a central role in shaping a unit's development and effectiveness. Nonetheless, the general insights from this section can be synthesised into a framework of design considerations consisting of three core parts: placement, practitioner heuristics and resilience mechanisms.

Placement links back to the OECD central steering, diffuse and project models. Decisions around placement are highly context-dependent, shaped by institutional structures, political conditions, underlying motivations and the maturity of behavioural policy within a given country.

Practitioner heuristics links directly to the APPLES framework as experienced-based rules of thumb that can help guide institutional design decisions for a nudge unit.

Resilience mechanisms capture the institutional features that enable a nudge unit to endure and adapt over time. Drawing on the BIG'R propositions, these include maintaining a stable team with effective handover mechanisms, developing an extensive and adaptable research toolkit and securing a sufficiently broad mandate to prevent the unit's scope from becoming overly constrained.

Policy Suggestions

1. Legally Anchored, Hybrid Institutional Design

The first policy recommendation concerns the institutional structuring of a nudge unit. At the core of this structuring, the unit should be legally anchored in terms of mandate, accountability and degree of independence. In practice, this could include a statute, formal inter-ministerial agreement or executive mandate, depending on the intended scope and permanency of the unit.

One effective way to structure the unit is through a hub-and-spoke model, where a central team provides methodological leadership, quality control and institutional memory, while behavioural specialists are embedded within line ministries to ensure sector-specific relevance and efficient implementation. This design effectively balances autonomy and coordination, and allows the institutional model to evolve over time.

2. Centralised Institutional Memory and Evidence Infrastructure

The second policy recommendation advocates that the unit must be established as a formal custodian of behavioural evidence. Under current governance, lessons from previous interventions are often lost within specific departments or project teams. However, if the unit is facilitated as a permanent repository for datasets, trial results, protocols and implementation notes, the cycle of "Test, Learn and Adapt" can be accelerated through agile governance and reduced costs. Furthermore, by documenting both the pre-intervention choice architecture and post-intervention impacts, an audit trail can be established to evaluate the validity of behavioural shifts and strengthen accountability.

3. Strategic Use of Nudges as Policy Complements, Not Substitutes

The third policy recommendation concerns the strategic positioning of behavioural interventions within the broader policy toolkit. Evidence from the Behavioural Insights Team's portfolio suggests that nudges are most effective as complements to—rather than substitutes for—structural, regulatory, or fiscal interventions.

The case studies show that nudges are particularly effective at optimising existing systems by reducing friction, increasing salience, or correcting attentional failures. Their wider significance, however, often lies in generating evidence that reveals suboptimal choice architectures and builds the case for more substantive reform. Nudges should therefore be understood not as end-point solutions, but as diagnostic tools that inform longer-term policy redesign and reduce uncertainty surrounding structural change.

Where possible, short-term nudges should also be paired with longer-term boosts that enhance citizens' decision-making capacities. This is particularly important in policy domains involving repeated decisions or long-run welfare effects, such as health, education, or financial planning.

4. Systematic Sludge Reduction as a Core Mandate

The fourth policy recommendation concerns reducing systematic sludge as a core mandate for the Nudge Unit. Excessive administrative friction, cognitive load, and bureaucratic hurdles can prevent people from accessing services or completing tasks. Current estimates suggest UK citizens spend roughly 3 billion hours annually on government-related admin, which means that tackling this burden could become central to government reform.

Focusing on high-burden public services, such as benefits applications, tax compliance, or licensing processes, it could become possible to generate large aggregate welfare gains by making small behavioural improvements. By implementing the EAST framework and using even relatively straightforward interventions such as creating default choices or removing redundant barriers, it may become possible to significantly increase the level of compliance with government-related admin requirements. This area offers high cost-effectiveness, strong ethical defensibility and relatively low political contestation.

5. Context-Sensitive Experimentation and Scaling Protocols

Failure at rollout can be due to context dependence. Behavioural effects are typically dependent on local norms, institutional frictions and baseline behaviour, resulting in different responses to the same intervention. Hence, scaling should not merely replicate a single successful pilot or lab experiment. Before a full-scale implementation, interventions must demonstrate robustness across diverse contexts. This includes (1) demographic variation (age, income, education, language), (2) geography (urban/rural, regions with different levels of service access), and (3) institutional settings (schools vs.\ clinics; digital vs.\ paper workflows).

Behavioural outcomes can also plausibly decay over time, due to habituation or implementation fatigue. Hence, short durations of performance tracking can systematically overstate long-run impact. Trials should therefore incorporate longer follow-up horizons commensurate with the policy objective.

Finally, the protocol should involve contextual reporting to a high level of fidelity so that others can replicate, diagnose failure, and adapt intelligently. Equipped with rich context-specific information, negative results in the future can be more interpretable. Policymakers would be better able to tell whether an intervention is ineffective or if it simply wasn't implemented under the conditions that generated the original effect.

6. Local Experimentation to Mitigate WEIRD Bias

The sixth policy recommendation concerns designing and running domestic experiments for behavioural policy, rather than importing policies wholesale from foreign contexts. While importing successful interventions from other countries may appear operationally cost-effective, such strategies are likely to overlook WEIRD bias and be scientifically reductive by ignoring local contexts. The evidence showed that statistical significance in foreign trials does not guarantee that effect sizes will generalise, and impacts may diminish or even reverse in diverse landscapes where motivational drivers differ.

Ultimately, governments should continue to invest in and expand the national nudge unit, transitioning it from a simple policy administrator to a research hub that conducts locally tailored experiments. In this way, universal principles can be refined and adapted to meet local needs and demands.

7. Ethical Targeting and Transparency Norms

The seventh policy recommendation concerns the ethical design, targeting, and governance of behavioural interventions. Behavioural policies should prioritise means-paternalistic interventions that assist individuals in achieving widely shared and publicly defensible ends, such as reducing administrative errors, improving comprehension, or facilitating access to public services.

In morally sensitive policy domains, behavioural interventions should favour deliberation-supporting designs over opaque or purely automatic defaults. Transparency should be institutionalised as a normative baseline rather than treated as a threat to behavioural effectiveness. The legitimacy of nudges should rest on clear justification, proportionality, and evidence of impact, not on invisibility.

To operationalise these principles, behavioural units should implement formal ethical review procedures at the design and pre-trial stage of interventions, assessed by an independent departmental ethics panel or cross-government behavioural oversight board, using structured criteria covering targeting, mechanism type, proportionality, and transparency.

8. Dedicated Funding Mechanism for Behavioural Trials

The eighth policy recommendation concerns ensuring effective behavioural policy through the design of funding arrangements for behavioural trials. Behavioural experimentation should be motivated by public value and learning, rather than by the likelihood of producing positive or easily demonstrable results. To facilitate this, nudge units should establish a central funding stream that departments can competitively access to run behavioural trials. Funding decisions should prioritise policy relevance, experimental rigour and learning potential, rather than expected success.

This funding model should be complemented by the mandatory dissemination of both positive and null results through a central evidence infrastructure, reinforcing a culture of experimentation and transparency rather than outcome cherry-picking. To ensure long-term

political buy-in, this approach should be complemented by clear performance metrics that demonstrate the unit's impact, emphasising long-term societal value rather than short-term outcomes.

VIII. Conclusion

This paper explored the impact of the UK's BIT on the growing use of behavioural insights in policy-making both domestically and internationally. It focused not only on the operational and policy-development aspects of the theme, but also on some controversial issues, such as the effectiveness of the nudge units and the ethical debates surrounding nudges.

The conceptual foundations of nudge were first clarified. Nudges were defined as light-touch behavioural interventions that predictably influence choices by redesigning the choice environment without bans, mandates or major incentive changes. The intellectual, ethical and practical roots of behavioural policymaking were traced through dual-systems thinking, libertarian paternalism and choice architecture, and behavioural economics was situated in relation to standard economic theory. Key mechanisms and taxonomies were outlined, alongside major critiques of behavioural approaches.

A comparative analysis of behavioural units across different political and cultural contexts then demonstrated that while many draw inspiration from the BIT template, institutional form varies significantly. Differences in evaluation methods, institutional location, staffing, authority and pathways to scale shape how units operate and how they define success. Political, economic and cultural contexts influence not only strategy, but also what counts as meaningful impact.

The paper further argued that a centralised national nudge unit can add value beyond isolated interventions. Institutionally, such a unit can enhance governance agility, preserve institutional memory, establish transparent standards and embed ethical oversight. Culturally, it can mitigate WEIRD evidence biases through local experimentation and context-sensitive design. It can also monitor spillover and distributional effects, strengthening long-term policy coherence, particularly in green and digital transitions.

Ethical debates were addressed directly. While critics raise concerns about autonomy, paternalism and manipulation, defenders emphasise the inevitability of choice architecture and the potential for autonomy-enhancing interventions. The analysis suggests that nudges should

neither be uncritically embraced nor abandoned, but applied with restraint, transparency and institutional safeguards.

The paper penultimately highlighted the central importance of institutional design. Drawing on the BIT experience, it proposed a framework of design considerations for behavioural units, informed by APPLES and BIG'R propositions and interpreted through a skunkworks lens. Rather than prescribing a single model, the framework underscores that effectiveness depends on alignment with institutional context, political incentives and administrative capacity. Behavioural governance, therefore, succeeds not through replication of a template, but through adaptive design grounded in local realities.

We hope that this paper will contribute meaningfully to ongoing debates among policymakers, fiscal institutions and research organisations about the appropriate role of behavioural insights in public governance. By outlining design principles for legally anchored, evidence-driven and ethically governed nudge units, we aim to contribute constructively to debates within HM Treasury, the Office for Budget Responsibility, European fiscal councils, and research organisations such as the Institute for Fiscal Studies, the Resolution Foundation and Bruegel, as well as their counterparts internationally. More broadly, we hope the framework offered here supports informed public dialogue by clarifying not only what behavioural interventions can achieve, but under what institutional conditions they are most likely to deliver durable, accountable and equitable policy outcomes.

Appendix

Appendix 1: Table displaying definitions of different nudge types extracted from Zhu (2024).

Nudge Type	Definition	Source
Default	Default nudges involve setting a pre-selected option that individuals automatically receive unless they actively choose an alternative. This leverages human inertia or status quo bias, making it easier for people to stick with the default choice rather than	Thaler \

	opting out. There are two main systems: opt-in (requiring individuals to actively choose to participate) and opt-out (automatically including individuals unless they choose not to participate).	
Feedback	Feedback nudges provide individuals with information about their past or present behaviours to influence future behaviour. It aims to help individuals understand the consequences of their previous or current actions and make adjustments. This nudge is commonly seen in interventions like energy consumption reports.	Thaler \
Framing	Framing nudges present information in a way that makes it either more attractive or less attractive. It leverages the framing effect, where decisions are influenced by how choices are presented. Common types of framing include loss-frame, gain-frame, altruistic-frame, and selfish-frame.	Tversky \
Incentives	Incentives are rewards or punishments that incentivise individuals to behave in certain ways. In the context of this research, incentives are categorised as a type of nudge, as most incentive	Lafont \

	nudges consist of small-scale rewards.	
Information	Information nudges involve presenting relevant data to raise awareness and enhance understanding. Key Point: Information disclosure does not consist of providing specific instructions that guide behaviours.	Corrales \
Instructive	Instructive nudges provide direct instructions to influence behaviour, often through polite prompts. These nudges leverage authority bias.	Braithwaite (2023)
Pre-commitment	Pre-commitment nudges involve an individual deciding in advance to bind oneself to a particular course of action.	Sunstein (2014); Dolan et al. (2010)
Priming	Priming nudges introduce one stimulus to influence responses to subsequent stimuli.	Dolan et al. (2010)
Reminder	Reminder nudges are prompts or cues designed to bring something to an individual's attention or to remind them to perform a specific action.	Damgaard \
Salience	Salience nudges make a particular choice or piece of information more prominent in the decision environment to increase uptake.	Bordalo, Gennaioli \

Simplification	Simplification nudges involve making information or processes easier to understand and use, reducing complexity to facilitate better decision-making.	Thaler \
Social norms	Social norms nudges use the accepted behaviours and beliefs within a setting to influence behaviour by showing what is typical or expected.	Dolan et al. (2010)

References

Barton, A. (2013). How tobacco health warnings can foster autonomy. *Public Health Ethics*, 6(2), 207–219.

Barton, A. and Grüne-Yanoff, T. (2015). From libertarian paternalism to nudging—and beyond. *Review of Philosophy and Psychology*, 6(3), 341–359.

Behavioural Insights Team. (2010). *Annual Update 2010–11.* https://casaa.org/wp-content/uploads/Behaviour-Change-Insight-Team-Annual-Update_acc.pdf

Behavioural Insights Team. (2011). *Annual Update 2011–12.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/83719/Behavioural-Insights-Team-Annual-Update-2011-12_0.pdf

Behavioural Insights Team. (2012). *Applying behavioural insights to reduce fraud, error and debt.* Cabinet Office. https://assets.publishing.service.gov.uk/media/5a789ae740f0b62b22cbb536/BIT_FraudErrorDebt_accessible.pdf

Behavioural Insights Team. (2012). *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials.* https://assets.publishing.service.gov.uk/media/5a7488c8e5274a7f9c586c23/TLA-1906126.pdf

Behavioural Insights Team. (2013). *Applying behavioural insights to organ donation: Preliminary results from a randomised controlled trial.* Cabinet Office.

https://assets.publishing.service.gov.uk/media/5a75ae4fe5274a436829925f/Applying_Behavioural_Insights_to_Organ_Donation.pdf

Behavioural Insights Team. (2014). *EAST: Four simple ways to apply behavioural insights*. https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights/

Behavioural Insights Team. (2017). *Increasing diversity in the police force*. https://www.bi.team/publications/levelling-the-playing-field-in-police-recruitment-evidence-from-a-field-experiment-on-test-performance/

Behavioural Insights Team. (2021). *Nudging to vaccinate: How BIT supported the COVID-19 response in Argentina*. https://www.bi.team/news/nudging-to-vaccinate-how-bit-supported-the-covid-19-response-in-argentina/

Behavioural Insights Team. (2023, August 9). Can nudging improve student wellbeing? Results from an RCT in Australia. https://www.bi.team/blogs/can-nudging-improve-student-wellbeing-results-from-an-rct-in-australia/

Behavioural Insights Team. (2024). *EAST: Four simple ways to apply behavioural insights* (Revised and updated ed.). https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights/

Behavioural Insights Team. (2025, April 9). 'Top of the queue' text led to around 42,000 additional vaccinations during Covid vaccine rollout [Press release]. https://www.bi.team/press-releases/top-of-the-queue-text-led-to-around-42000-additional-vaccinations-during-covid-vaccine-rollout/

Behavioural Insights Team. (2025). Our history. https://www.bi.team/about-us/our-history/

Blair, T. (2010). *A Journey*. London: Hutchinson.

Blumenthal-Barby, J. S. and Burroughs, H. (2012). Seeking better health care outcomes: The ethics of using the "nudge". *The American Journal of Bioethics*, 12(2), 1–10.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3), 1243–1285. https://doi.org/10.1093/qje/qjs018

Brown, D., Barrera, A., Ibañez, L. et al. (2024). A behaviourally informed chatbot increases vaccination rates in Argentina more than a one-way reminder. *Nature Human Behaviour*, 8, 2314–2321. https://doi.org/10.1038/s41562-024-01985-7

Caldwell, L. (2018). Public and private sector nudgers can learn from each other. *Behavioural Public Policy*, 2(2). https://www.cambridge.org/core/journals/behavioural-public-policy/article/public-and-private-sector-nudgers-can-learn-from-each-other/AB1FD60FA0FA73BB747A9C66A6C44212

Cialdini, R. B. (2007). *Influence: The psychology of persuasion* (Rev. ed.). Harper Business.

Conly, S. (2013). *Against autonomy: Justifying coercive paternalism*. Cambridge: Cambridge University Press.

Damgaard, M. T. and Gravert, C. (2017). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics*, 157, 15–26.

Dewies, M., Merkelbach, I., Edelenbos, J., Rohde, K. I. M. and Denkt\c{a}ş, S. (2023). Comprehensive Evaluation of the Behavioural Insights Group Rotterdam. *Administration and Society*, 55(8), 1555–1583. https://pure.eur.nl/ws/files/155472874/dewies-et-al-2023-comprehensive-evaluation-of-the-behavioral-insights-group-rotterdam.pdf

Dolan, P., Hallsworth, M., Halpern, D., King, D. and Vlaev, I. (2010). *MINDSPACE: Influencing behaviour through public policy*. Institute for Government & Cabinet Office. https://www.instituteforgovernment.org.uk/sites/default/files/publications/MINDSPACE.pdf

Engelen, B. and Nys, T. (2019). Nudging and autonomy: Analyzing and alleviating the worries. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-019-00450-z

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181. https://doi.org/10.1257/jep.14.3.159

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5), 589–611. https://doi.org/10.1111/1467-6419.00150

Ghesla, C., Gsottbauer, A. and Schubert, R. (2020). Nudging the poor and the rich—A field study on the distributional effects of green electricity defaults. *Energy Economics*, 86, 104616. https://doi.org/10.1016/j.eneco.2019.104616

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S. and Xu, Y. (2019). Measuring individual differences in effort: Evidence from a series of social experiments. https://doi.org/10.1257/aeri.20180633

Halpern, D. (2015). *Inside the nudge unit: How small changes can make a big difference*. WH Allen.

Halpern, D. and Sanders, M. (2016). Nudging by government: Progress, impact, & lessons learned. *Behavioural Science & Policy*, 2(2), 52–65. https://doi.org/10.1353/bsp.2016.0015

Hausman, D. M. and Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, 18(1), 123–136.

Haynes, L., Service, O., Goldacre, B. and Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. Cabinet Office. https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials

Henrich, J., Heine, S. J. and Norenzayan, A. (2010). The weirdest people in the world? *Behavioural and Brain Sciences*, 33(2–3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Hernández-Agramonte, J. M. and Espinoza Iglesias, K. (2023). MineduLab, the innovation laboratory for a cost-effective educational policy in Peru. In M. Sanders, S. Bhanot and S.

O'Flaherty (Eds.), *Behavioural Public Policy in a Global Context: Practical Lessons from Outside the Nudge Unit*. Palgrave Macmillan.

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

House, J., Lacetera, N., Macis, M. and Mazar, N. (2024). Nudging the nudger: Performance feedback and organ donor registrations. *Journal of Health Economics*, 97, 102914. https://doi.org/10.1016/j.jhealeco.2024.102914

John, P. (2014). Policy entrepreneurship in UK central government: The behavioural insights team and the use of randomised controlled trials. *Public Policy and Administration*, 29(3). https://doi.org/10.1177/0952076713509297

Johnson, E. J. and Goldstein, D. (2003). Do defaults save lives? *Science*, 302(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Jones, S., Head, B. and Ferguson, M. (2021). In search of policy innovation: Behavioural Insights Teams in Australia and New Zealand. *Australian Journal of Public Administration*, 80(3). https://doi.org/10.1111/1467-8500.12478

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

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. https://doi.org/10.2307/1914185

Koch, C., Nafziger, J. and Nielsen, H. S. (2023). Spillover effects of reminder nudges in complex environments. *Proceedings of the National Academy of Sciences*, 120(43). https://doi.org/10.1073/pnas.2322549121

Komatsu, S., Kaneko, S. and Fujii, H. (2022). Searching for the universality of nudging: A cross-cultural comparison of the information effects of reminding people about familial support. *PLOS ONE*, 17(11), e0277969. https://doi.org/10.1371/journal.pone.0277969

Kuehnhanns, C. R. (2019). The challenges of behavioural insights for effective policy design. *Policy and Society*, 38(1), 14–40. https://doi.org/10.1080/14494035.2018.1511188

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2), 443–477. https://doi.org/10.1162/003355397555253

Leggett, W. (2014). The politics of behaviour change: Nudge, neoliberalism and the state. *Policy & Politics*, 42(1), 3–19. https://doi.org/10.1332/030557312X655576

MacKay, D. and Robinson, A. (2016). The ethics of organ donor registration policies: Nudges and respect for autonomy. *The American Journal of Bioethics*, 16(11), 3–12.

Martinez, S.-K., Meier, S. and Sprenger, C. (2022). Procrastination in the field: Evidence from tax filing. *Journal of the European Economic Association*. https://doi.org/10.1093/jeea/jvac067

Mendizabal, E. (2024). Beyond the hype: Lessons and challenges from MineduLab's quest for evidence-informed policy-making. *On Think Tanks*. https://onthinktanks.org/articles/beyond-the-hype-lessons-and-challenges-from-minedulabs-quest-for-evidence-based-policy-making/

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

Mills, S. (2025). How "nudge" happened: The political economy of nudging in the UK. *Cambridge Journal of Economics*, 39(1), 1–18. https://doi.org/10.1093/cje/beu040

Mullainathan, S. and Shafir, E. (2013). *Scarcity: Why having too little means so much*. Henry Holt and Company.

Nesta. (2021). Nesta acquires Behavioural Insights Team to help tackle UK's biggest social challenges. https://www.nesta.org.uk/press-release/nesta-acquires-behavioural-insights-team/

Neuhaus, T. and Curley, L. (2022). The emergence of global behavioural public policy—developments of and within the nudge unit. *World Complexity Science Academy Journal*, 3(2), 1–17. https://www.researchgate.net/publication/364335324

OECD. (2017). *Behavioural insights and public policy: Lessons from around the world*. OECD Publishing. https://doi.org/10.1787/9789264270480-en

OECD. (2024). *LOGIC: Good practice principles for mainstreaming behavioural public policy*. OECD Publishing. https://doi.org/10.1787/6cb52de2-en

Oliver, A. (2013). From nudging to budging: Using behavioural economics to inform public sector policy. *Journal of Social Policy*, 42(4), 685–700. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/D98361CED793BE761AA22BF49299BF43/S0047279413000299a.pdf

Quinn, B. (2018, November 10). The 'nudge unit': the experts that became a prime UK export. *The Guardian*. https://www.theguardian.com/politics/2018/nov/10/nudge-unit-pushed-way-private-sector-behavioural-insights-team

Ruda, S. (2022). Will nudge theory survive the pandemic? *UnHerd*, 13 January. https://unherd.com/2022/01/how-the-government-abused-nudge-theory/

Sallis, A., Harper, H. and Sanders, M. (2018). Effect of persuasive messages on National Health Service Organ Donor Registrations: a pragmatic quasi-randomised controlled trial with one million UK road taxpayers. *Trials*, 19(1), 1–10. https://doi.org/10.1186/s13063-018-2855-5

Samuelson, W. and Zeckhauser, R. J. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59. https://doi.org/10.1007/BF00055564

Sanders, M., Snijders, V. and Hallsworth, M. (2018). Behavioural science and policy: Where are we now and where are we going? *Behavioural Public Policy*, 2(2), 144–167.

Schmidt, A. T. (2017). The power to nudge. *American Political Science Review*, 111(2), 404–417.

Schmidt, A. T. and Engelen, B. (2020). The ethics of nudging: An overview. *Wiley Interdisciplinary Reviews*, 1–13.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J. and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5), 429–434. https://doi.org/10.1111/j.1467-9280.2007.01917.x

Service, O. (2014, March 21). The Behavioural Insights Team and randomised controlled trials. *Nesta*. https://www.nesta.org.uk/blog/the-behavioural-insights-team-and-randomised-controlled-trials/

Siipi, H. (2025). Danger of slippery slopes in nudge research. *Journal of Academic Ethics*, 23, 695–715.

Simon, H. A. (1955). A behavioural model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118. https://doi.org/10.2307/1884852

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Sunstein, C. R. (2016). *The ethics of influence: Government in the age of behavioural science*. Cambridge University Press.

Sunstein, C. R. (2022). The distributional effects of nudges. *Nature Human Behaviour*, 6(1), 9–10. https://doi.org/10.1038/s41562-021-01236-z

Talhelm, T. (2025). Adapting interventions to culture can improve effectiveness and cost-efficiency. *Policy Insights from the Behavioural and Brain Sciences*, 12(1). https://doi.org/10.1177/23727322251408076

Thaler, R. H. and Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, 93(2), 175–179. https://doi.org/10.1257/000282803321947001

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

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

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wintour, P. (2014, February 5). Government's behaviour insight team to become a mutual and sell services. *The Guardian*. https://www.theguardian.com/politics/2014/feb/05/government-behaviour-insight-nudge-mutual-nesta-funding

Wright, O. (2013). How organ donation is getting nudge in the right direction. *The Independent*, 24 December. https://www.independent.co.uk/news/uk/politics/how-organ-donation-is-getting-nudge-in-the-right-direction-trial-could-pave-way-for-100-000-extra-donors-each-year-9023349.html

Zey, M. (2001). Rational choice and organisation theory. *International Encyclopedia of the Social & Behavioural Sciences*, 12751–12755. https://doi.org/10.1016/b0-08-043076-7/04212-1

Zhu, X. (2024). *An examination of nudge interventions in Japan and its implications for Japanese behavioural public administration*. University of Tokyo, Graduate School of Public Policy. https://www.pp.u-tokyo.ac.jp/wp-content/uploads/2016/02/An-Examination-of-Nudge-Interventions-in-Japan-and-its-Implications-for-Japanese-Behavioural-Public-Administration.pdf