Trends in Cognitive Sciences

CellPress
REVIEWS

Opinion

Learning from Behavioural Changes That Fail

Magda Osman,[1,*] Scott McLachlan,[2,3] Norman Fenton,[2] Martin Neil,[2] Ragnar Löfstedt,[4] and Björn Meder[5,6,*]

Behavioural change techniques are currently used by many global organisations and public institutions. The amassing evidence base is used to answer practical and scientific questions regarding what cognitive, affective, and environment factors lead to successful behavioural change in the laboratory and in the field. In this piece we show that there is also value to examining interventions that inadvertently fail in achieving their desired behavioural change (e.g., backfiring effects). We identify the underlying causal pathways that characterise different types of failure, and show how a taxonomy of causal interactions that result in failure exposes new insights that can advance theory and practice.

Highlights

The behavioural change enterprise disproportionately focuses on promoting successes at the expense of examining the failures of behavioural change interventions.

We review the literature across different fields through a causal explanatory approach to identify structural relations that impede (or promote) the success of interventions.

Based on this analysis we present a taxonomy of failures of behavioural change that catalogues different types of failures and backfiring effects.

Our analyses and classification offer guidance for practitioners and researchers alike, and provide critical insights for establishing a more robust foundation for evidence-based policy.

Behavioural Change: What Is It, Who Uses it, and Why?

Using psychological insights to motivate people to change their opinions, attitudes, and behaviour go back as far as Ancient Greece [1] and Rome [2]. The 20th century sparked a dedicated interest in the psychological processes underlying **behavioural change** (see Glossary) from psychotherapists [3] and psycholinguists [4] studying the art of persuasion through to the behavioural engineering enterprise of the behaviourists in the 1950s [5] and 1960s [6], and on to nudges in the 2000s [7,8]. Early in history, as today, the behavioural change enterprise has been designed to answer a practical problem: what levers can we pull to change behaviour reliably both for personal and public good? Today, many public and private institutions implement **behavioural change frameworks** and **behavioural change theories**, which aim to provide leverage for changing behaviour on a mass scale (Box 1) [9]. Essentially, the goal is to identify and describe the factors that determine a targeted behaviour, decompose the contexts in which it is observed, and design and implement interventions to induce positive change. Applications range from efforts to improve dietary choices [10] and help people save more for their retirement [11] through to communicating responses to a public health crisis [12].

Learning from Failure

How successful are behavioural interventions? [13]. The Organisation for Economic Co-operation and Development (OECD) created a catalogue of interventions and treatments (we use these terms interchangeably) implemented in public bodies in 23 countries, showcasing the success story of behavioural change around the world (Box 2).The problem is that a rather loose inclusion criterion was used to determine what counts as an intervention (e.g., literature review), raising concerns about what is deemed a success [14]. Conversely, while failures have recently received some attention in the literature [15–17], along with illustrations of particular types (e.g., **backfire effects**) [18], a systematic evaluation of failures is lacking. In the researcher's world, this might be the result of publication bias; this has been instrumental in the replication crisis in psychology [19]. In the practitioner's world, the preoccupation with the credibility of evidence to determine the efficacy of a treatment overshadows examining what has not worked and why. In fact, this is where a **causal explanatory approach** would benefit both the basic science and policy world: constructing causal scenarios of potential relations that distinguish what might be relevant (what features of the intervention you think would influence an outcome) from what is not relevant (extraneous

[1]Biological and Experimental Psychology Group, Queen Mary University of London, London, UK
[2]Risk and Information Management, Queen Mary University of London, London, UK
[3]Health Informatics and Knowledge Engineering Research (HiKER) Group, Queen Mary University of London, London, UK
[4]Department of Geography, Kings College London, London, UK
[5]Faculty of Health, Health and Medical University, Potsdam, Germany
[6]Max Planck Institute for Human Development, Berlin, Germany

*Correspondence:
m.osman@qmul.ac.uk (M. Osman) and
meder@mpib-berlin.mpg.de (B. Meder).

Box 1. Behaviour Change Frameworks: Wheels, Checklists, Diamonds, and Ladders

One way of appreciating the popularity of the behavioural change enterprise is the sheer volume of frameworks and theories of behavioural change that have been developed – by recent counts there are at least 83 such frameworks [68,69]. These frameworks are used by several institutions, including OECD, World Bank, and the World Health Organization (WHO), and governmental agencies in several countries worldwide. The purpose is to provide practical and ethical guidance for designing and implementing behaviour change techniques [65,66].

Behaviour Change Wheel (BCW)

The framework has two purposes: it is a model of behaviour, as well as a way of designing interventions [70]. The BCW has embedded in it the COM-B (Capability, Opportunity, Motivation, and Behaviour) system. Regardless of whether selected interventions are based on existing evidence or not, the BCW and COM-B explicitly assert the value, and necessity, of theory-driven predictions.

MINDSPACE (Messenger, Incentives, Norms, Defaults, Salience, Priming, Affect, Commitment, and Ego)

MINDSPACE is a checklist of nine key influences that psychological and economic research have shown to influence behaviour, within narrow as well as broader environments [71,72]. The establishment of criteria for success are stringent. The practitioner must state in advance the direction, magnitude, and period of change in the outcome following the intervention. This approach has been refined by the UK Behavioural Insight Team to four basic principles: Easy, Attractive, Social, and Timely (EAST) [73].

BASIC (Behaviours, Analysis, Strategies, Interventions, and Change)

BASIC is a behavioural change toolkit developed and used by the OECD [74,75]. The toolkit is prescriptive. Practitioners have several different charts, and diagrams, and checklists that help definine the policy problem, behavioural outcomes, interventions, and the ethicality of the processes, all of which are needed to state in advance the rationale for the intervention and the expected outcomes.

Nuffield Ladder of Intervention

The Nuffield Council on Bioethics ladder of intervention was designed in the context of public health but since has been applied to many other domains outside of health [76]. The focus is on ethical issues surrounding behaviour change techniques. As interventions become more intrusive with respect to the autonomy of individuals, communities, or the population, the greater the effort needed to justify in advance the predicted outcomes, relative to the overall public good.

Nudge

Rather than a framework, nudge is a collection of approaches sharing certain characteristics (e.g., changing behaviour without much leverage) to alter choice environments (so-called choice architectures) to achieve behaviour change [17,18]. Its definition is prescriptive because it asserts that the success of a nudge is gauged by the predicted outcome. The nudge approach is rooted in the heuristics-and-biases program [77,78] and dual-process frameworks of cognition [79].

factors that influence an outcome), from which to reason effectively over what has been observed and why [20].

In our view, the upshot from all this is an underappreciation of when and why behaviour change fails, and the lack of a clear conceptualisation of different types of failure. The present article (i) highlights that reports of failure and backfiring are common in the literature, across a wide array of fields and types of interventions; (ii) identifies characteristic regularities and causal pathways underlying these failures; and (iii) presents a taxonomy derived from the identified commonalities, which can be applied by researchers and practitioners to better characterise and analyse different types of failure and map out the broader factors in which interventions are trialled.

Crucially, the proposed taxonomy goes beyond the widespread analytic perspective that behavioural trials mainly fail because they do not exert any influence on the target behaviour, or the inability of effects to scale up, or to persist in the long run. Rather, our classification illustrates that there are several causal scenarios and conditions that can lead to different kinds of failure. Thus, our taxonomy should help practitioners to better anticipate possible failures of a planned intervention, gauge the likelihood of failure for a given context and policy problem, and improve

Glossary

Backfire effect: outcomes in choices and actions resulting from treatments trialled by researchers and practitioners of behavioural change that are in opposite direction to those predicted.
Behavioural change: use of interventions informed by behavioural and social science disciplines that are designed for the purpose of supporting decisions and actions that lead to short term and long-term changes that benefit the individual and society.
Behavioural change framework: synthesis of theoretical constructs that are organised in such a way to help researchers and practitioners identify the cognitive, affective, social, and environmental (economic, physical) influences on behaviour, but that does not explicitly propose testable relationships between these constructs.
Behavioural change theory: description of the psychological mechanisms that characterise human cognition and behaviour such that they enable researchers and practitioners to generate testable predictions of the effects of treatments (moderators of change in behaviour) on behaviour.
Boomerang effects: as a consequence of a treatment, typically reported with reference to social norms and often reported in social psychology, whereby an individual will interpret the message negatively, or interpret the message as intended but deliberately fail to comply with the message.
Causal explanatory approach: process of characterising the structural relationships between variables of interest so that it is possible to infer the influence of data across a causal network to assess the influences of interventions on outcomes.
Reactance: where an experienced threat (whether real or perceived) to personal freedom and control leads to changes in the valuation of the threated choice such that motivations increase to restore the choice, either by acting counter to the threat, withdrawing actions altogether, or introducing a new action in the threatened context.
Rebound effects: typically observed in domains where treatments are designed to increase energy efficiency, the unintended and unanticipated outcome is that potential energy savings induce new behavioural responses that later undercut the original savings observed.

Box 2. Future Proofing Behavioural Change in Social and Public Policy Making

There are many innovations that are being explored by governments and intergovernmental organisations (e.g., OECD and WHO) to improve behavioural change approaches for addressing social and public policy needs. The motivation is to share best practices as well as developing tool kits to improve understanding of the policy needs to match behavioural solutions to.

An illustration of a significant innovation, which is in line with open science initiatives in the cognitive and other sciences, is the development of a database by the OECD [9,80]. They promoted the need to develop a catalogue of behavioural interventions that have been implemented by government organisations. They led by example, and were one of the first to compile their own catalogue of 111 behavioural change interventions implemented in 23 countries, ranging widely in the policy sectors they were targeting (i.e., business, charity, consumer protection, education, energy, environment, financial products, health and safety, labour market, public service delivery, tax, and telecommunications). Later work examining the OECD's catalogue of behavioural interventions showed that it was lacking in basic reporting of key methods (e.g., sample sizes and number of treatment conditions) and statistical details (e.g., effect sizes and statistical analyses) [14].

Nonetheless, this is still a step in the right direction because there is a need for a database that is regularly updated with interventions that have been trialled in OECD countries, and beyond. It would a vehicle for making data sets publicly accessible, allowing researchers and practitioners to ask questions and evaluate the strength of the available evidence. This will ultimately lead to a better understanding of what interventions work where and what does not work, with a view to also improving understanding of why they do work, and why they do not work. This same approach could be implemented in government departments or even a cross-governmental database, so that sharing of findings can be achieved more efficiently to expose the successes as well as failures of trialled interventions.

In clinical medicine there exists a Template for Intervention Description and Replication (TIDieR) which is an agreed approach for ensuring the accuracy of reporting of interventions, and a means of standardising reporting to aid replicability [81]. This, along with the cost associated with implementing the trials, would allow for a thorough assessment of the value for money of the potential gains from successful interventions against the cost of their implementation, as well as the costs of different ranges of failures, or inadvertent successes.

Spillovers: because of treatments trialled by researchers and practitioners of behavioural change the observed behavioural change in one context can either be amplified, eliminated, or reversed in another context or related behaviours within the original context.

trial design and, ultimately, policy. From a theoretical perspective, such analyses are critical for determining the strengths and limits of current behavioural change frameworks, and help to put behavioural change on a robust theoretical foundation [21,22] (J. Hastings *et al.*, unpublished). Importantly, our causal approach is not restricted to qualitative causal analyses but also provides an array of computational methods for quantitative causal modelling of behaviour change and public policy (Box 3) [24–26].

Examining Failures of Behavioural Change

How often do behaviour change interventions fail? What is clear from reviewing the literature is that failed interventions are surprisingly common, that their prevalence presents looming consequences for the field of behavioural interventions, and that not all failures are the same. To gauge the frequency and type of failures we began with search terms 'behavioural change', 'behavioural intervention', 'nudge', and 'nudging', entered into Web of Science and Google Scholar between 2008 to 2019. We filtered the obtained literature to include only articles that referred to terms backfire, **spillovers**, **rebounds**, and **boomerang effects**, and excluded all review articles. For the full details of the filtering process, as well as a comprehensive table that catalogues the 65 studies according to several details (e.g., date, author, country, study type, design, intervention, summary of findings, and conclusions) see https://psyarxiv.com/ae756.

Of the 65 studies that were compiled, 58% included a field experiment, and 75% of all studies also included a control or baseline condition to compare the behavioural interventions against. Common domains in which the interventions were trialled included charitable donations (13%), tax compliance (8%), health (diet or exercise; 25%), and proenvironmental behaviour (28%). Sometimes multiple interventions were included in a single study, so we report here the intervention that was identified by the researchers of the study as critical to achieving

Box 3. Bayesian Causal Analytic Approach

We show the value of examining policy situations through a causal analytic lens. To illustrate, Figure I presents a simplified causal Bayesian network model [24–26] to characterise the probabilistic relationships between relevant variables in a scenario for which behavioural interventions are used to increase actual donations of organs. Even without any formal knowledge, drawing graphical causal models can help make inferences about where and how to intervene with a behavioural intervention, as well as deriving precise and empirically testable predictions [82–85].

So, how does one build a causal model? It contains three basic elements: nodes, arrows, and probabilities. Nodes represent the domain variables (with either two or more states) and their associated numerical value (probabilities associated with each state). The arrows indicate presumed probabilistic causal relationships between variables. The strength of the relations is captured in probability tables (not shown in Figure I) whose values are either learned from data or elicited from domain experts. To start with, expert judgment is needed to interpolate probabilities, because there are policy domains where observed specific values for variables may not yet exist, but models can be easily updated with new data to increase the accuracy of the estimates of the effects of behavioural interventions.

The causal model shown in Figure IA comprises four nodes with binary states. The outcome of interest is the likelihood of organ donation in a single case, and the model presents arrows between the default policy in place (opt-in, opt-out), the rates at which families will not consent (veto) organ donation of their dying relative, and the rates at which people are on the donation register.

We consider two scenarios presented in Figure IB (opt-in) and Figure IC (opt-out); this is where we fix the policy variable and let the model reveal the effect on the outcome. It is likely that knowing whether a dying relative made an active decision (opt-in) to donate carries a clearer signal of their underlying wishes, as compared to a passive decision (opt-out) [86]. This impacts the extent to which a family considers the policy on which decisions to donate are expressed. Hence, although there is a large increase in registered donors (60% for opt-out compared to 20% for opt-in) there is also a proportionally much higher probability of veto. So, the overall effect is only a small increase in actual organ donations (from 14% to 16%).

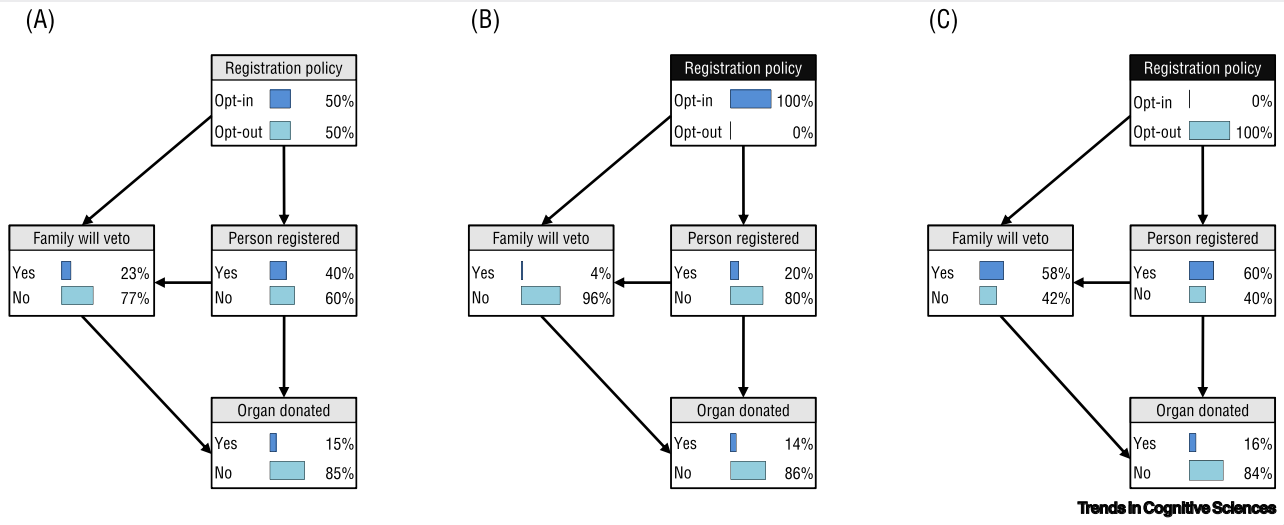

Figure I. Simplified Causal Bayesian Network Models Representing the Probabilistic Relationships between Variables in the Domain of Organ Donation. (A) The causal network without any evidence entered, showing the probability distribution of each variable. (B) Setting the policy variable to a default opt-in system results in updated probability values for all variables, based on the presumed relations. Under an opt-in policy, both registration rates and the likelihood of the family vetoing are rather low, and so is the probability of an organ actually being donated. (C) Under an opt-out default, registration rates are much higher but also families are more likely to veto. Consequently, only a small increase in donation rates results, compared with an opt-in scenario.

behavioural change. The 65 studies utilised several types of interventions, notably defaults (15%), social comparisons and social norming (40%), labelling (12%), and provision of information delivered through letters or text messaging (24%). Below we discuss in detail several key examples of different types of failures, with Table 1 summarising core details of these illustrations.

Taking a Causal Approach to Characterising Failed Interventions

Our review of the literature shows that failures vary from anything between the predicted outcome not being produced or even generating the opposite outcome, to generating unintended side

Table 1. Illustrative Examples of the Types of Failures Presented in the Taxonomy of Failures of Behavioural Change in Figure 1[a]

Type of failure	Type of study	Design	Domain	Behavioural change intervention	Refs
(A) No treatment effect	Field	Randomised control trial (RCT)	Environmental sustainability	Social comparison; provision of information, delivered by weekly mobile text messages	[30]
	Field	RCT	Tax compliance	Saliency manipulation, messaging, delivered through letter	[31]
(B) Backfiring	Laboratory study 2	RCT	Health (eating)	Framing of dietary options	[34]
	Field	Pre- vs post-interventions	Health (beverages) and environmental sustainability	Restructuring the choice set, bottled water ban; provision of information, delivered through an educational campaign	[36]
(C) Treatment offset by negative side effect	Laboratory study 1– 6	RCT	Environmental sustainability; Saving for retirement	Defaults; framing; carbon tax	[38]
	Field	RCT	Health (eating)	Saliency manipulation; provision of information, delivered via a menu	[41]
(D) No treatment effect, but positive side effect	Laboratory study 2	RCT	Health (eating)	Saliency manipulation; sizing of packaging	[87]
	Laboratory study 1	RCT	Environmental sustainability; Charitable donation	Saliency manipulation; restructuring the choice set	[88]
(E) Only proxy changes, no actual criterion	Field	Observational	Organ donation	Defaults	[42]
	Laboratory	RCT	Health (eating)	Tailored information; provision of information, delivered as a document; decision time	[46]
(F) Treatment offset by later behaviour	Field	RCT	Tax compliance	Injunctive social norm; framing; provision of information, delivered in a tax bill	[89]
	Field	RCT	Charitable donations	Saliency manipulation; message reminders, delivered as emails	[50]
(G) Environment does not support changes	Field	Observational	Organ donation	Defaults; provision of information, delivered via public campaigns	[55]
	Field	RCT	Environmental sustainability	Injunctive social norms; labelling, delivered via signage in supermarket isles	[90]
(H) Intervention triggers counteracting forces	Laboratory	RCT	Personal finances	Saliency manipulation; messaging, delivered as instructional information	[58]
	Field	Pre- vs post-interventions	Health (beverages)	Sizing of packaging; public campaigns; sugar tax	[59]

[a]For full set of studies, see https://psyarxiv.com/ae756.

effects that offset the desired behaviour change. The question is, how best to characterise the different failures? Figure 1, Key Figure provides an overview of different types of failures, each described using simple qualitative causal models to distinguish between different scenarios and outcomes (see later for details). Distinguishing between different types of failures and the corresponding causal scenarios is critical in evidence-based policy contexts [20,27,28], other practical contexts [26], and basic science contexts [29]. Why might this be useful? Essentially when planning an intervention, one needs to ask oneself: what factors could be causally relevant to the success of the intervention? How could the intervention influence these factors? What precautionary measures should be taken to avoid failure? Without addressing these questions early on, researchers risk trialling interventions that may work but cannot be scaled up to a wider population. Also, the interventions may not preserve their effects over time. This is because the underlying mechanisms or other relevant factors are not understood because they may compete with or undo the change achieved earlier. Without an understanding of the reasons for their failure, interventions present costs, both in time and public funds.

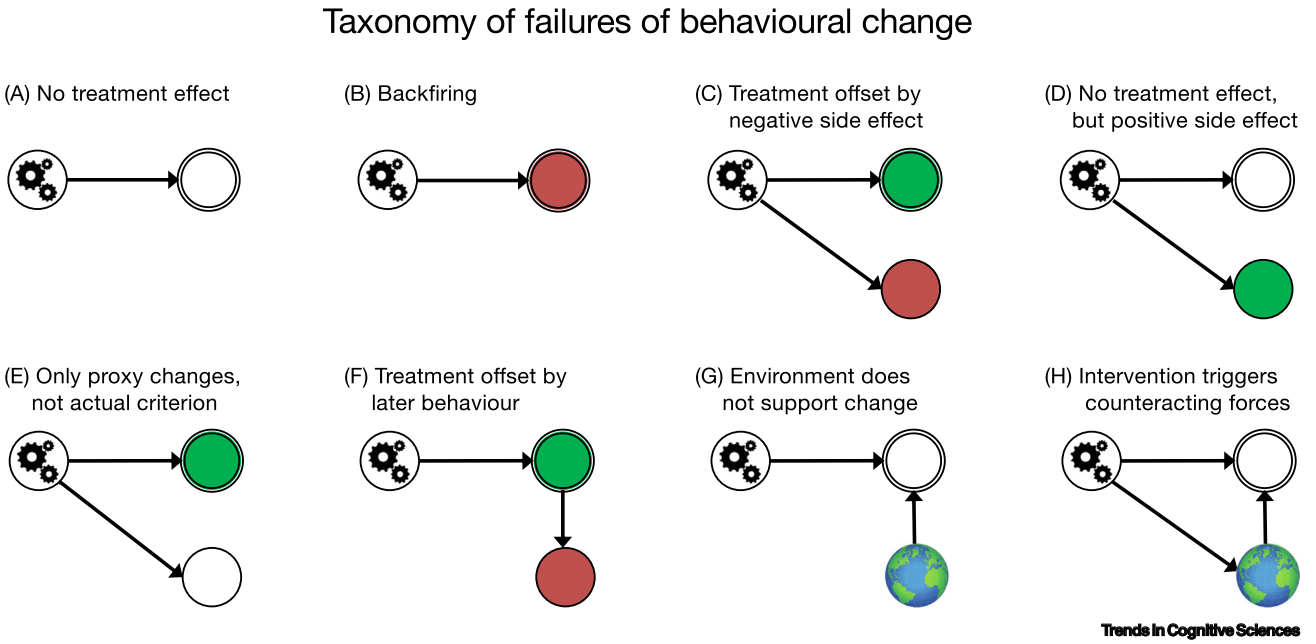

Figure 1. Key Figure. Each type of failure is represented in the form of a simple qualitative causal model, where the arrows depict presumed causal relations among the domain variables (nodes). The gears represent the intervention and the node with the double outline representing the target behaviour (e.g., energy consumption and calorie intake). Green and red code for positive and negative outcomes, respectively, and these are relative to the goal of the intervention and a baseline. White indicates that the intervention failed to influence the targeted behaviour. The additional node in (C–F) represents further domain variables or behaviours that were not intended or foreseen to be influenced by the intervention (e.g., positive or negative side effects). The globe in (G) and (H) represents the broader environment which may have direct unforeseen influences on the outcome and can also be influenced by the treatment (e.g., environment does not support the intended behavioural change or counteracting forces in the environment offset the effect).

This is where causal counterfactuals prove to be valuable as a way of scientifically thinking through hypotheticals and different causal scenarios by providing a source of practical gain for researchers and practitioners alike. The ultimate value of taking this kind of approach is that it encourages thinking in a more principled way about the causal structure of the domain and the policy problem. Rather than asking in hindsight after an intervention failed 'what went wrong?', researchers and practitioners should ask in advance 'what could go wrong, and how could it go wrong?'. Our main point, which is reiterated by others, is that anticipating possible types of failures and potential causal pathways in the initial stages of designing interventions could, and should, guide the development of behavioural interventions in a systematic fashion. Doing this exposes the relevant factors and potential causal dependencies beyond the local cause and effect relationships that are at the heart of the policy problem. Conducting such a principled analysis in advance is particularly helpful when conducting field experiments, where there is great difficulty in controlling for externalities.

Types of Failure

No Treatment Effects

The most basic type of failure are no treatment effects, where the intervention was unable to result in behaviour change relative to baseline (Figure 1A). Using a social comparison nudge to reduce water consumption may fail to do so overall, and even lead subgroups to increase their water consumption [30]. Tax authorities may use a deterrent or moral persuasion message to increase tax compliance and find that this has no overall effect, with some corporations even incurring higher tax deductions [31]. Using financial incentives to increase people's physical activity might fail to induce behavioural change [32], and providing calorie information may fail to

decrease calorie intake [33]. Such findings are helpful to determine which behavioural interventions do not work, or work only under specific circumstances, or for particular subgroups, thereby helping practitioners to focus their efforts on the most promising interventions.

Backfiring

The prototypical case of backfiring occurs when the intervention does change the target behaviour but in the opposite direction to what was intended (Figure 1B). For instance, many educational campaigns aim to change dietary choices by providing information on the negative consequences of unhealthy food consumption. In practice, this does not always pan out as planned. People with strong concerns regarding their weight and physical fitness (dieters) can express higher desire for and show increased consumption of unhealthy foods after receiving a message highlighting the negative aspects of particular food items, or a lack of exercise [34]; an example of **reactance**. Anti-soda advertisements used in public campaigns can in fact increase, rather than decrease, consumption of sugary drinks, both in the laboratory and the field [23,35]. There are also cases of multibackfiring, where an intervention yields negative effects on different outcomes (e.g., increased consumption of sugar-sweetened beverages and failure to reduce the number of disposable bottles) [36]. Careful analysis of such occurrences can also help practitioners to avoid interventions with diametral effects. For instance, using descriptive norms, which communicate typical behaviours of a target group (rather than injunctive norms which communicate approved or disapproved behaviours), can increase undesirable behaviours [37].

Treatment Offset by Negative Side Effects

The profiles of treatment offset by negative side effects are different from backfires because positive treatment effects are observed but are offset by negative side effects, such that the positive change is attenuated or eliminated (Figure 1C). A green energy default nudge can decrease support for more comprehensive (but also more onerous) policies like a carbon tax [38]. An environmental campaign may reduce residents' water consumption but at the same time increase their electricity consumption [39]. Introducing a tax on saturated fats can decrease fat intake and increase vegetable and fruit consumption but also increase salt consumption [40]. In situations of this kind, treatment effects can be offset because of compensatory choices, for instance, information regarding calories can increase the selection of healthier options but is offset by higher calorific sides and beverages [41]. Failures like this highlight the importance of considering multiple outcome criteria to enable a comprehensive evaluation of interventions along with the possible compensatory behaviours that can later undo the intervention.

No Treatment Effect, but Positive Side Effects

In no treatment effect, but positive side effects examples, interventions yield unforeseen positive consequences, even if the actual target behaviour remains unchanged (Figure 1D). For instance, a comparison of countries with different default policies for organ donation (opt-in vs opt-out system) shows no difference in overall transplant rates, thereby raising concerns about the effectiveness of implementing an opt-out strategy to increase organ donation rates [42]. At the same time, however, more fine-grained analyses reveal diverging effects when considering live and deceased donor rates separately. For example, while the total number of kidney transplants does not vary between opt-in and opt-out countries (i.e., no overall treatment effect), opt-out countries have a higher number of deceased donors and a lower number of living donors, compared with opt-in countries. This constitutes a positive side effect, because fewer live donors are subject to immediate risk of harm from organ harvesting and associated long-term risks and possible reductions in life expectancy [43]. Generally, assessing positive side effects is important, because it is often hypothesised (tacitly or explicitly) that changing specific behaviours (e.g., reducing water consumption or healthier lunches) will generalise to other behaviours in

related contexts as well (e.g., increase environmentally friendly behaviours such as recycling could in turn increase consumption of sustainable foods) [44,45].

Only Proxy Changes, not Actual Criterion

An important policy goal is to influence behaviour at the population level, and this can leave practitioners liable to only proxy changes, not actual criterion; this can be for pragmatic reasons, or because criterion data is difficult to obtain. Changes by proxy essentially refer to behavioural changes that are pragmatic substitutes for the target behaviour (e.g., becoming a potential organ donor – which is a proxy for actual organs donated). However, changes on surrogate criteria do not always translate to changes in key target criteria (Figure 1E). A higher number of potential organ donors in countries with an opt-out system does not necessarily translate to a higher number of organ donations, but instead can have diametric consequences when focusing specifically on different kinds of organ donations (e.g., higher deceased donor rates but lower living donor rates) [42]. Similarly, information provision may increase healthy food selections in a simulated supermarket but have no long-term impact on body mass index and lifestyle [46]. Such cases illustrate the importance of assessing the sustainability of interventions in the long run and with respect to relevant target criteria, regardless of the additional difficulty that comes with assessing success.

Positive Treatment Effect Is Offset by Later Behaviour

Even if an intervention successfully alters a target behaviour, it can happen that a positive treatment effect is offset by later behaviour (Figure 1F) [47,48]. Encouraging office occupants to adopt environmentally friendly behaviours by reducing the default temperature on their thermostats can reduce energy consumption for small reductions in temperature – but also make people over-ride the default when reductions are too large, thereby undercutting the treatment effect [49]. Charitable organisations sending reminders to potential donors can increase donations, but later lead to higher unsubscriber rates from the mailing list, thereby jeopardising future donations [50]. Taking dietary supplements can increase their consumption but reduce people's desire to engage in physical exercise [51]. These findings are critical because they point to key psychological processes that need to be considered in behaviour change such as moral licensing [52], where a virtuous decision can later lead to indulgent behaviour [53].

Environment Does not Support Change

Because of the complexities of introducing interventions in live settings, what can happen is that the environment does not support change. In such cases it only becomes clear with hindsight that the effect of an intervention critically depends on the broader environment to support behavioural change, but if the required affordances are not present the intervention will fail to achieve its goals (Figure 1G) [54]. For instance, in the 1980s, Croatia moved from an opt-in to an opt-out system for organ donations, but for several years this had little impact on actual donation rates because the necessary medical infrastructure was not in place (e.g., insufficient numbers of staff and equipment needed to carry out transplant operations) [55]. Introducing bike-sharing systems to increase cycling and reduce traffic congestions and pollution fail if a major concern of users is road safety [56,57]. What these examples highlight is that behaviour change interventions cannot be implemented at the exclusion of acknowledging factors in the broader environment that are critical to facilitating the intended change.

Intervention Triggers Counteracting Forces

There are also cases where the intervention triggers counteracting forces, in which case positive effects are counteracted by forces in the broader environment (which is different to simply not having the available affordance to support behavioural change) (Figure 1H). A case in point is the use of salience messaging to help curb unnecessary spending, where the general

dispositions of subsamples (one group were 'spendthrifts' extravagant with money, the other group were 'tightwads' miserly with money) counteract the overall impact of the treatment on behaviour [58]. Financial institutions can actively counteract policy changes designed to protect consumers. For instance, introducing defaults to limit access to overdrafts of bank accounts, but then include structures that enable the defaults to be overridden easily [17]. Consumers may struggle to curb their consumption of unhealthy sugary beverages in response to regulators' choice restricting methods (e.g., sugar tax) because drinks companies challenge the regulation [59]. Examples of this kind highlight context-specific externalities that reside in the environment and the various ways in which they can counteract interventions.

Concluding Remarks and Future Directions

The current appetite for using behaviour change techniques is undeniable, with terms such as nudge having become part of scientific and public vernacular. But where do failures fit into the behavioural change enterprise?

Before addressing this question, we need to draw attention to a critical limit in the conclusions we can draw. In talking about what status failures should have, we need to recognise that we simply do not (yet) have a precise idea of the scale and range of failed interventions. While our survey of the literature shows that failures are not isolated cases, our analyses are not a reliable quantitative gauge to assess the extent to which the problem is systemic. This points to a problem regarding a lack of interest in, or even a systemic failure of reporting failed interventions. If the latter, this may be because of publication bias [19]. If the former, then one reason might be that there is not an agreed scientific language or terminology for classifying different failures, for which we now present a taxonomy that can help with this, and that serves as a starting point for further and more definitive analyses of failures (see Outstanding Questions).

There is a corollary to this, of course. We also do not have a comprehensive and nuanced way of reporting the outcomes of interventions from which we can gain a comprehensive idea of their scale, range, and robustness. The field of behavioural change would benefit from being more accepting of different levels and types of success and failure, as well as more systematic classifications of the strength of the available evidence. In this sense, behaviour change research faces similar challenges as other fields that are at the intersection of basic and applied science. Institutions like the Cochrane Collaboration (https://www.cochrane.org/) in medicine and What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/) in education provide practitioners with detailed and systematic reviews that synthesise and evaluate the existing literature, to enable robust evidence-based recommendations and provide information tailored to practitioners' needs and decision-making processes. Establishing similar initiatives in the domain of behaviour change would be valuable from both an applied and basic science view (Box 2). Building an evidence hierarchy which classifies and evaluates interventions on scales of successes and failures would enable an accurate depiction of the outcomes and which populations they impact to different degrees [60,61]. This is all the more important because the literature on behaviour change now spans across many disciplines, problem domains, and methodologies, raising the need to keep track of the enterprise to ensure that basic research and application mutually inform each other.

Future Empirical Work

Notwithstanding the important caveat that we do not have an appropriate base rate of failure and in which social policy domains they are most likely to occur, the most common intervention used that resulted in failures were social norming or social comparisons. Why is this? It could be that both happen to be used more frequently than other interventions because they are cheap,

Outstanding Questions

How common are behavioural change intervention failures and is there publication bias? Which interventions fail more often, and in which domains do they most commonly fail? Can a dedicated database be generated and maintained that classifies the different types of failures in ways that can be used by researchers and practitioners?

In the future, should failures receive the same amount of attention as successes? If yes, how do researchers and practitioner ensure this so to avoid potential problems with how failures are reported?

Given that a behavioural intervention can both succeed and fail to different degrees, can we develop a system by which we can score interventions on separate scales of success and failure to obtain a more comprehensive representation of the behavioural outcomes they generate to expand our conceptualisation of them?

What are common cognitive and behavioural signatures of failed interventions that help to (in)validate cognitive and social theories of behaviour?

How do we know when we have achieved successful behavioural change and what are the most relevant indicators of success? Is it the magnitude of positive change, the extent of positive spillovers, the level of uptake in the target population, or the longevity of the behavioural change?

Can computational modelling approaches (e.g., causal Bayesian network approaches) be used to map the causal structure of a context where behavioural change techniques are implemented, to derive precise predictions regarding success and failure?

Should behavioural change frameworks, and more broadly cognitive and psychological theories, be modified to consider the dynamics of behaviour to identify key factors that motivate change in cognitions, emotions, and behaviour over time?

What methods can be used to determine in a principled manner, in

simple to implement, and potentially scalable. Their ubiquity bears out in our findings since context does not seem to be predictive of failure. We find them applied in such varied domains, including reduction of energy and water consumption, as well as to increase exercise, healthy eating, tax compliance, and charitable donations. Another possible explanation for their frequent failures is that not all social norms are created equally [62], posing diametral consequences for their effectiveness in some but not other types [37]. However, the divergent influence of different kinds of social norms observed in the laboratory has not consistently been obtained in large-scale field studies, where different subgroups of a target population have been found to respond differently to social norming messages. For instance, in the domain of energy conversation, the effectiveness of a social comparison nudge providing information on a given household's consumption relative to other households can depend on their overall consumption [63] or recipients' political ideology [64].

advance, the length of time a behavioural change technique should be trialled to ensure that it has been given the best chance of not failing (or best chance of succeeding)?

In fact, a recurring theme across many studies and types of failures is that subgroups matter. Thus, to the extent that subgroups are part of the manipulation, or demographic details enable *post hoc* analyses, we see that for different subgroups, the interventions either worked, did not take, or backfired. This constitutes a practical gain for systematically analysing failures and can help answer the question 'what works for whom and why (not)?' by pointing researchers and practitioners to relevant factors they need to consider before embarking on new studies. These, in turn, can be guided by a causal analysis where these factors are explicitly represented to make predictions about the effectiveness of an intervention under different scenarios (e.g., varying proportions of political ideology in the target group). Thus, failures serve as a valuable resource for future interventions. In the absence of a systematic analysis and record of failures, there is no way of anticipating where interventions might fail, in order to improve trial design to surmount the failure, or trial a different intervention that is more likely to succeed.

Theory Development

How best can advances be achieved in the theoretical and methodological foundations of behaviour change research? What is common to almost all behaviour change frameworks (Box 1) is the absence of a causal model. The exception is the Behaviour Change Wheel, which uses the COM-B (Capability, Opportunity, Motivation, and Behaviour) system as a causal explanatory approach of behaviour. However, given the causal structure of the model, it is ill equipped to support a formal causal analysis that can meanfully interpret the direction of effect from any intervention on any one component. Computational causal modelling techniques can complement existing behaviour change frameworks by offering an account to formally model behaviour change problems and the context in which interventions are situated. From this perspective, the focus is not on different theoretical perspectives and assumptions about psychological processes [22], but on building robust and predictive models of behaviour change interventions that explicate and formally represent the assumed causal dependencies. Box 3 illustrates this kind of analysis and causal model building for the domain of organ donation, using the framework of causal Bayesian networks [24–26,29]. By mapping the presumed relations and incorporating the relevant uncertainties this approach can bridge different theoretical perspectives and enable researchers to derive model-based predictions for the effects of interventions. In this case the status of failures is illustrative of the value of a causal analytic approach that should be adopted to improve theorising and hypothesis testing [91]. This analysis can be accompanied by framing efforts to support behavioural change by focusing on boosting people's own competencies (e.g., improving risk literacy) and abilities to help them to exercise their own agency [65,66]. As Samuel Beckett said: 'Ever tried. Ever failed. No Matter. Try again. Fail again. Fail better' [67].

Acknowledgements

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under project EP/P009964/1: PAMBAYESIAN: Patient managed decision support using Bayesian Networks. This work was also supported in part by the Economic and Social Research Council (ESRC) under project ES/R00787X/1: Rebuilding Macroeconomics, Putting in effort for the benefit of all: the role of reward and effort requirements.

References

1. Dimitrev, S. (2011) *The Greek Slogan of Freedom and Early Roman Politics in Greece*, Oxford University Press
2. Evans, J.D. (1992) *The Art of Persuasion: Political Propaganda from Aeneas to Brutus*, University of Michigan Press
3. Williams, T.A. (1909) The difference between suggestion and persuasion: the importance of the distinction. *Alienist Neurol.* 2, 1–10
4. Woolbert, C.H. (1917) Conviction and persuasion: some considerations of theory. *Q. J. Speech* 3, 249–264
5. Breland, K. and Breland, M. (1951) A field of applied animal psychology. *Am. Psychol.* 6, 202–204
6. Homme, L. *et al.* (1968) What behavioral engineering is. *Psychol. Rec.* 18, 425–434
7. Thaler, R.H. and Sunstein, C.R. (2008) *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press
8. Thaler, R.H. (2016) Behavioral economics: past, present, and future. *Am. Econ. Rev.* 106, 1577–1600
9. Organisation for Economic Co-operation and Development (OECD) (2017) *Behavioural Insights and Public Policy Lessons from around the World*, OECD Publishing
10. Hollands, G.J. *et al.* (2013) Altering micro-environments to change population health behaviour: towards an evidence base for choice architecture interventions. *BMC Public Health* 13, 1218
11. Benartzi, S. and Thaler, R.H. (2013) Behavioral economics and the retirement savings crisis. *Science* 339, 1152–1153
12. Van Bavel, J.J. *et al.* (2020) Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* 4, 460–471
13. Hummel, D. and Maedche, A. (2019) How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *J. Behav. Exp. Econ.* 80, 47–58
14. Osman, M. *et al.* (2020) Learning lessons: how to practice nudging around the world. *J. Risk Res.* 23, 11–19
15. West, R. *et al.* (2020) *Achieving Behaviour Change: a Guide for National Government*, Public Health England
16. Sunstein, C. (2017) Nudges that fail. *Behav. Public Policy* 1, 4–25
17. Willis, L.E. (2013) When nudges fail: slippery defaults. *Univ. Chic. Law Rev.* 80, 1155–1229
18. Stibe, A. and Cugelman, B. (2016) Persuasive backfiring: when behavior change interventions trigger unintended negative outcomes. In *International Conference on Persuasive Technology*, pp. 65–77, Springer
19. Tincani, M. and Travers, J. (2019) Replication research, publication bias, and applied behavior analysis. *Perspect. Behav. Sci.* 42, 59–75
20. Cartwright, N. (2009) Evidence-based policy: what's to be done about relevance? *Philos. Stud.* 143, 127–136
21. Muthukrishna, M. and Henrich, J. (2019) A problem in theory. *Nat. Hum. Behav.* 3, 221–229
22. Hastings, J. *et al.* (2020) Theory and ontology in behavioural science. *Nat. Hum. Behav.* 4, 226
23. DeSantis, J.A. (2011) Formulating a soda tax fit for consumption: a pragmatic approach to implementing the failed New York soda tax. *Mich. St. UJ Med. & L.* 16, 363
24. Spirtes, P. *et al.* (2000) *Causation, Prediction, and Search*, MIT Press
25. Pearl, J. (2000) *Causality: Models, Reasoning and Inference*, Cambridge University Press
26. Fenton, N. and Neil, M. (2012) *Risk Assessment and Decision Analysis with Bayesian Networks*, CRC Press
27. Cartwright, N. and Hardie, J. (2012) *Evidence-Based Policy: a Practical Guide to Doing It Better*, Oxford University Press
28. Cartwright, N. (2012) RCTs, evidence and predicting policy effectiveness. In *The Oxford Handbook of Philosophy of Social Science*, pp. 298–318, Oxford University Press
29. Pearl, J. and Mackenzie, D. (2018) *The Book of Why: the New Science of Cause and Effect*, Basic Books
30. Chabe-Ferret, S. *et al.* (2019) Can we nudge farmers into saving water? Evidence from a randomised experiment. *Eur. Rev. Agric. Econ.* 46, 393–416
31. Ariel, B. (2012) Deterrence and moral persuasion effects on corporate tax compliance: findings from a randomized controlled trial. *Criminology* 50, 27–69
32. Hunter, R.F. *et al.* (2013) Physical activity loyalty cards for behavior change: a quasi-experimental study. *Am. J. Prev. Med.* 45, 56–63
33. Swartz, J.J. *et al.* (2011) Calorie menu labeling on quick-service restaurant menus: an updated systematic review of the literature. *Int. J. Behav. Nutr. Phys. Act.* 8, 135
34. Pham, N. *et al.* (2016) Messages from the food police: how food-related warnings backfire among dieters. *J. Assoc. Consum. Res.* 1, 175–190
35. Debnam, J. and Just, D.R. (2017) *Endogenous responses to paternalism: examining psychological reactance in the lab and the field*. files.webservices.illinois.edu/7370/jakinadebnamjmp.pdf
36. Berman, E.R. and Johnson, R.K. (2015) The unintended consequences of changes in beverage options and the removal of bottled water on a university campus. *Am. J. Public Health* 105, 1404–1408
37. Schultz, P.W. *et al.* (2007) The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* 18, 429–434
38. Hagmann, D. *et al.* (2019) Nudging out support for a carbon tax. *Nat. Clim. Chang.* 9, 484–489
39. Tiefenbeck, V. *et al.* (2013) For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign. *Energy Policy* 57, 160–171
40. Smed, S. *et al.* (2016) The effects of the Danish saturated fat tax on food and nutrient intake and modelled health outcomes: an econometric and comparative risk assessment evaluation. *Eur. J. Clin. Nutr.* 70, 681–686
41. Wisdom, J. *et al.* (2010) Promoting healthy choices: information versus convenience. *Am. Econ. J.* 2, 164–178
42. Arshad, A. *et al.* (2019) Comparison of organ donation and transplantation rates between opt-out and opt-in systems. *Kidney Int.* 95, 1453–1460
43. Kiberd, B.A. (2013) Estimating the long term impact of kidney donation on life expectancy and end stage renal disease. *Transp. Res.* 2, 2
44. Department for Environment, Food and Rural Affairs (DEFRA) (2008) *A Framework for Pro-environmental Behaviours*, DEFRA
45. Thøgersen, J. and Crompton, T. (2009) Simple and painless? The limitations of spillover in environmental campaigning. *J. Consum. Policy* 32, 141–163
46. Belot, M. *et al.* (2020) Facilitating healthy dietary habits: an experiment with a low income population. *Eur. Econ. Rev.* 129, 103550
47. Dolan, P. and Galizzi, M.M. (2015) Like ripples on a pond: behavioral spillovers and their implications for research and policy. *J. Econ. Psychol.* 47, 1–16
48. Truelove, H.B. *et al.* (2014) Positive and negative spillover of pro-environmental behavior: an integrative review and theoretical framework. *Glob. Environ. Chang.* 29, 127–138
49. Brown, Z. *et al.* (2013) Testing the effect of defaults on the thermostat settings of OECD employees. *Energy Econ.* 39, 128–134
50. Damgaard, M.T. and Gravert, C. (2018) The hidden costs of nudging: experimental evidence from reminders in fundraising. *J. Public Econ.* 157, 15–26

51. Chiou, W.B. *et al.* (2011) Ironic effects of dietary supplementation: illusory invulnerability created by taking dietary supplements licenses health-risk behaviors. *Psychol. Sci.* 22, 1081–1086
52. Blanken, I. *et al.* (2015) A meta-analytic review of moral licensing. *Personal. Soc. Psychol. Bull.* 41, 540–558
53. Khan, U. and Dhar, R. (2006) Licensing effect in consumer choice. *J. Mark. Res.* 43, 259–266
54. Meder, B. *et al.* (2018) Beyond the confines of choice architecture: a critical analysis. *J. Econ. Psychol.* 68, 36–44
55. Živčić-Ćosić, S. *et al.* (2013) Development of the Croatian model of organ donation and transplantation. *Croat. Med. J.* 54, 65–70
56. Pucher, J. and Buehler, R. (2016) Safer cycling through improved infrastructure. *Am. J. Public Health* 106, 2089–2091
57. Pucher, J. and Dijkstra, L. (2003) Promoting safe walking and cycling to improve public health: Lessons from the Netherlands and Germany. *Am. J. Public Health* 93, 1509–1516
58. Thunström, L. *et al.* (2018) Nudges that hurt those already hurting–distributional and unintended effects of salience nudges. *J. Econ. Behav. Organ.* 153, 267–282
59. Fairchild, A.L. (2013) Half empty or half full? New York's soda rule in historical perspective. *N. Engl. J. Med.* 368, 1765–1767
60. Vaismoradi, M. *et al.* (2013) Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nurs. Health Sci.* 15, 398–405
61. Joffe, H. and Yardley, L. (2004) Content and thematic analysis. In *Research Methods for Clinical and Health Psychology* (Marks, D.F. and Yardley, L., eds), pp. 56–68, Sage
62. Cialdini, R.B. (2003) Crafting normative messages to protect the environment. *Curr. Dir. Psychol. Sci.* 12, 105–109
63. Allcott, H. (2011) Social norms and energy conservation. *J. Public Econ.* 95, 1082–1095
64. Costa, D.L. and Kahn, M.E. (2013) Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *J. Eur. Econ. Assoc.* 11, 680–702
65. Hertwig, R. and Grüne-Yanoff, T. (2017) Nudging and boosting: steering or empowering good decisions. *Perspect. Psychol. Sci.* 12, 973–986
66. Hertwig, R. (2017) When to consider boosting: some rules for policy-makers. *Behav. Public Policy* 1, 143–161
67. Beckett, Samuel (1983) *Worstword Ho*, John Calder
68. Davis, R. *et al.* (2015) Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychol. Rev.* 9, 323–344
69. West, R. *et al.* (2019) Development of a formal system for representing behaviour-change theories. *Nat. Hum. Behav.* 3, 526–536
70. Michie, S. *et al.* (2011) The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement. Sci.* 6, 42
71. Dolan, P. *et al.* (2012) Influencing behaviour: the mindspace way. *J. Econ. Psychol.* 33, 264–277
72. Institute for Government (2010) *MINDSPACE: Influencing Behaviour through Public Policy*, Institute for Government, the Cabinet Office
73. Team, B.I. (2014) *EAST: Four Simple Ways to Apply Behavioural Insights*, Behavioural Insight Team
74. Hansen, P.G. and Jespersen, A.M. (2013) Nudge and the manipulation of choice: a framework for the responsible use of the nudge approach to behaviour change in public policy. *Eur. J. Risk Regul.* 4, 3–28
75. Hansen, P.G. (2019) *Tools and Ethics for Applied Behavioural Insights: the BASIC Toolkit*, Organisation for Economic Co-operation and Development
76. Nuffield Bioethics Council (2007) *Public Health: Ethical Issues*, Nuffield Council on Bioethics
77. Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131
78. Gigerenzer, G. (2015) On the supposed evidence for libertarian paternalism. *Rev. Philos. Psychol.* 6, 361–383
79. Kahneman, D. (2011) *Thinking, Fast and Slow*, Macmillan
80. Organisation for Economic Co-operation and Development (OECD) (2017) *Tackling Environmental Problems with the Help of Behavioural Insights*, OECD
81. Hoffmann, T.C. *et al.* (2014) Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 348, g1687
82. Mohammadfam, I. *et al.* (2017) Constructing a Bayesian network model for improving safety behavior of employees at workplaces. *Appl. Ergon.* 58, 35–47
83. Xu, S. *et al.* (2018) Modeling interrelationships between health behaviors in overweight breast cancer survivors: Applying Bayesian networks. *PLOS ONE* 13, e0202923
84. Buck, C. *et al.* (2019) Factors influencing sedentary behaviour: a system based analysis using Bayesian networks within DEDIPAC. *PLOS ONE* 14, e0211546
85. Cook, J. and Lewandowsky, S. (2016) Rational irrationality: modeling climate change belief polarization using Bayesian networks. *Top. Cogn. Sci.* 8, 160–179
86. Lin, Y. *et al.* (2018) Underlying wishes and nudged choices. *J. Exp. Psychol. Appl.* 24, 459–475
87. Coelho do Vale, R. *et al.* (2008) Flying under the radar: perverse package size effects on consumption self-regulation. *J. Consum. Res.* 35, 380–390
88. Margetts, E.A. and Kashima, Y. (2017) Spillover between pro-environmental behaviours: the role of resources and perceived similarity. *J. Environ. Psychol.* 49, 30–42
89. Castro, L. and Scartascini, C. (2015) Tax compliance and enforcement in the pampas evidence from a field experiment. *J. Econ. Behav. Organ.* 116, 65–82
90. Richter, I. *et al.* (2018) A social norms intervention going wrong: Boomerang effects from descriptive norms information. *Sustainability* 10, 2848
91. Grune-Yanoff, T.Marchionni, C.Feufel, M.A. (2018) Toward a framework for selecting behavioral policies: how to choose between boosts and nudges. *Econ. Philos.* 34 (2), 243–266