Trends in Cognitive Sciences

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 Meder5,6,*

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.

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 factors leading 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.

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 factors leading 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.
Thus, our taxonomy should help practitioners to better anticipate possible failures of a planned intervention, 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 understanding of the 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 highlights that reports of failure and backfiring are common in the literature, across a wide array of fields and types of interventions; identifies characteristic regularities and causal pathways underlying these failures; and 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 understanding of the factors that influence an outcome, from which to reason effectively over what has been observed and why [20].

In view of 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 highlights that reports of failure and backfiring are common in the literature, across a wide array of fields and types of interventions; identifies characteristic regularities and causal pathways underlying these failures; and 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 understanding of the factors that influence an outcome, from which to reason effectively over what has been observed and why [20].
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 catalogue 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 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].

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.
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 effects. 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.
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.

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

<table>
<thead>
<tr>
<th>Type of failure</th>
<th>Type of study</th>
<th>Design</th>
<th>Domain</th>
<th>Behaviour change intervention</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) No treatment effect</td>
<td>Field</td>
<td>Randomised control trial (RCT)</td>
<td>Environmental sustainability</td>
<td>Social comparison; provision of information, delivered by weekly mobile text messages</td>
<td>[30]</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>RCT</td>
<td>Tax compliance</td>
<td>Saliency manipulation, messaging, delivered through letter</td>
<td>[31]</td>
</tr>
<tr>
<td>(B) Backfiring</td>
<td>Laboratory study 2</td>
<td>RCT</td>
<td>Health (eating)</td>
<td>Framing of dietary options</td>
<td>[34]</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>Pre- vs post-interventions</td>
<td>Health (beverages) and environmental sustainability</td>
<td>Restructuring the choice set, bottled water ban; provision of information, delivered through an educational campaign</td>
<td>[36]</td>
</tr>
<tr>
<td>(C) Treatment offset by negative side effect</td>
<td>Laboratory study 1–6</td>
<td>RCT</td>
<td>Environmental sustainability; Saving for retirement</td>
<td>Defaults; framing; carbon tax</td>
<td>[38]</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>RCT</td>
<td>Health (eating)</td>
<td>Saliency manipulation; provision of information, delivered via a menu</td>
<td>[41]</td>
</tr>
<tr>
<td>(D) No treatment effect, but positive side effect</td>
<td>Laboratory study 2</td>
<td>RCT</td>
<td>Health (eating)</td>
<td>Saliency manipulation; sizing of packaging</td>
<td>[67]</td>
</tr>
<tr>
<td></td>
<td>Laboratory study 1</td>
<td>RCT</td>
<td>Environmental sustainability; Charitable donation</td>
<td>Saliency manipulation; restructuring the choice set</td>
<td>[68]</td>
</tr>
<tr>
<td>(E) Only proxy changes, no actual criterion</td>
<td>Field</td>
<td>Observational</td>
<td>Organ donation</td>
<td>Defaults</td>
<td>[42]</td>
</tr>
<tr>
<td></td>
<td>Laboratory</td>
<td>RCT</td>
<td>Health (eating)</td>
<td>Tailored information; provision of information, delivered as a document; decision time</td>
<td>[46]</td>
</tr>
<tr>
<td>(F) Treatment offset by later behaviour</td>
<td>Field</td>
<td>RCT</td>
<td>Tax compliance</td>
<td>Injunctive social norm; framing; provision of information, delivered in a tax bill</td>
<td>[89]</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>RCT</td>
<td>Charitable donations</td>
<td>Saliency manipulation; message reminders, delivered as emails</td>
<td>[50]</td>
</tr>
<tr>
<td>(G) Environment does not support changes</td>
<td>Field</td>
<td>Observational</td>
<td>Organ donation</td>
<td>Defaults; provision of information, delivered via public campaigns</td>
<td>[55]</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>RCT</td>
<td>Environmental sustainability</td>
<td>Injunctive social norms; labelling, delivered via signage in supermarket isles</td>
<td>[90]</td>
</tr>
<tr>
<td>(H) Intervention triggers countering forces</td>
<td>Laboratory</td>
<td>RCT</td>
<td>Personal finances</td>
<td>Saliency manipulation; messaging, delivered as instructional information</td>
<td>[58]</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>Pre- vs post-interventions</td>
<td>Health (beverages)</td>
<td>Sizing of packaging; public campaigns; sugar tax</td>
<td>[59]</td>
</tr>
</tbody>
</table>

*For full set of studies, see https://psyarxiv.com/ae756.
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 caloric 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 unsubscribe 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,
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 diametrical 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].

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 meaningfully 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: PAWBYESEAN: 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

3. Williams, T.A. (1909) The difference between suggestion and persuasion: the importance of the distinction. Alienist Neuro. 2, 1–10
5. Breland, K. and Breland, M. (1951) A field of applied animal psychology. Am. Psychol. 6, 202–204
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
73. Team, B.J. (2014) EAST: Four Simple Ways to Apply Behavioural Insights, Behavioural Insight Team