Abstract

A nudge is a paternalistic government intervention that attempts to improve choices by changing the framing of a decision problem. We propose a welfare-theoretic foundation for nudging similar in spirit to the classical revealed preference approach, by investigating a framework where preferences and mistakes of an agent can be elicited from her choices under different frames. We provide characterizations of the classes of behavioral models in which the information required for nudging can or cannot be deduced from choice data.

*Keywords*: nudge, framing, behavioral welfare economics, revealed preference

**JEL Classification**: D03, D04, D60, D82

---

‡We are grateful for comments by Sandro Ambühl, Sandeep Baliga, Eddie Dekel, Kfir Eliaz, Samuel Häfner, Igor Letina, Konrad Mierendorff, Tim Netzer, Georg Nöldeke, Ariel Rubinstein, Yuval Salant, Armin Schnitzler, Ron Siegel, Ran Spiegler, Georg Weizsäcker, seminar audiences at DICE Düsseldorf, European University Institute, Goethe University Frankfurt, HECER Helsinki, Johannes Gutenberg University Mainz, Northwestern University, NYU Abu Dhabi, Tel Aviv University, UCL, the Universities of Basel, Bern, Bonn, Konstanz, Michigan, Surrey, and Zurich, and participants at CESifo Area Conference on Behavioural Economics 2014, BBE Workshop 2015, Midwest Economic Theory Meeting Fall 2015, Swiss Economists Abroad Meeting 2015, Verein für Socialpolitik Theoretischer Ausschuss 2015, and BERA Micro Workshop 2016. All errors are our own.

∗University of Zurich, Department of Economics, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland, and UBS International Center of Economics in Society at the University of Zurich. Email: jean-michel.benkert@econ.uzh.ch. The author would like to thank the UBS International Center of Economics in Society at the University of Zurich and the Swiss National Science Foundation (Doc.Mobility Grant P1ZHP1_161810) for financial support.

∗∗University of Zurich, Department of Economics, Bluemlisalpstrasse 10, 8006 Zurich, Switzerland. Email: nick.netzer@econ.uzh.ch.
1 Introduction

A nudge (Thaler and Sunstein, 2008) is a regulatory intervention that is characterized by two properties. First, it is paternalistic in nature, because “it is selected with the goal of influencing the choices of affected parties in a way that will make those parties better off” (Thaler and Sunstein, 2003, p. 175). Second, it is not coercive but instead manipulates the framing of a decision problem. Among the best-known examples already discussed in Thaler and Sunstein (2003) is retirement saving in 401(k) savings plans, which can be encouraged tremendously by setting the default to automatic enrollment. Another example is the order in which food is presented in a cafeteria, which can be used to promote a more healthy diet.

The intriguing idea that choices can be improved by framing has made nudging also politically attractive. Governments of numerous countries have set up so-called “nudge units,” which develop and implement nudge-based policies. The UK spearheaded this development in 2010 with the foundation of the Behavioral Insights Team.1 More recently, former US President Barack Obama issued an executive order that encourages all government agencies to “carefully consider how the presentation and structure of [...] choices, including the order, number, and arrangement of options, can most effectively promote public welfare”.2

The definition of welfare remains an intricate problem, because the behavioral incoherence caused by framing renders an ordinary revealed preference approach infeasible. The applied nudging literature often takes as given a behavior that the regulator wants to implement (for in-depth discussions see Grüne-Yanoff, 2012; Goldin, 2015). There are indeed applications in which the goal of nudging is fairly uncontroversial. For example, introspection may already tell us that smokers appreciate (or at least do not oppose) soft measures which help them smoke less, like a less prominent display of tobacco products in stores. Such an intervention can also be motivated by the fact that smoking causes negative externalities on third parties. The nudge is not paternalistic in that case, but simply a novel policy instrument for which a fundamental welfare problem does not arise (see e.g. arguments in Chetty et al., 2009; Mullainathan et al., 2012; Handel, 2013; Allcott and Taubinsky, 2015; Baicker et al., 2015).

Prior knowledge of the desired behavior is clearly sufficient for evaluating the success of a nudge. In this paper, we ask whether prior knowledge is necessary. Do we need to know ex ante that the employee should save more or eat more healthily? Or can the relevant information about welfare be inferred from observed choices, in the same way that welfare is not predetermined but inferred from choices in the classical revealed preference approach? There are two reasons for asking this question. First, in interesting

---

1See http://www.behaviouralinsights.co.uk.
2See http://go.wh.gov/MKURtv.
applications the desired behavior is presumably not known ex ante. The optimal trade-off between current and future consumption is not obvious, and the same holds for health and the pleasure of unhealthy food. Second, on a more fundamental level, the non-coercive nature of nudging is particularly compelling when there is a reason for granting freedom of choice, on top of an assumed intrinsic value. If the regulator knows the behavior that she wants to induce, she could dictate that behavior right away. For instance, negative externalities are considered a sufficient reason for smoking bans in public areas. A forceful argument in favor of nudging would show that the regulator can learn enough to pick an optimal framing, without already knowing the behavior that this framing will induce.

We provide an answer to this question in a framework where the welfare preference of an agent may be inferred from her choices under different frames. This is done by explicitly modelling the distortionary effects of frames. We thus follow the model-based approach to behavioral welfare economics. Our answer is multifaceted. There are behavioral models for which learning about an optimal nudge is possible and relatively easy. There are also models for which learning is possible but very complex. Finally, there are models for which learning the relevant information is fundamentally impossible.

Our formal framework builds on Rubinstein and Salant (2012), henceforth RS, who formulate a generalized approach for eliciting an agent’s preferences from choice data. In this framework, which we formally introduce in Section 2, a regulator has a conjecture about the behavioral model \( d \), which relates each pair of a welfare preference \( \succeq \) and a frame \( f \) to a behavioral preference \( d(\succeq, f) \). The interpretation is that an agent with welfare preference \( \succeq \) acts as if maximizing \( d(\succeq, f) \) if the situation is framed by \( f \). The welfare preference represents the normatively relevant well-being of the agent but is not observable. Behavioral preferences are in principle observable in the usual revealed preference sense, but they may be different from the welfare preference. RS investigate the problem of learning about the welfare preference from a data set that contains observations of behavior. We are interested in evaluating the frames based on the acquired information about welfare.

The formal framework assumes that, holding fixed a frame, choices are coherent across choice sets and can therefore be represented by a behavioral preference. Incoherence arises when the framing is changed. Salant and Rubinstein (2008) refer to choice functions with this property as “salient consideration functions,” and Ariely et al. (2003) use the term “coherent arbitrariness.” It rules out behaviors which violate standard axioms already when a frame is fixed, such as choice with limited attention in the formulation of Masatli-

\[3\]Kőszegi and Rabin (2008b) first emphasized the possibility of recovering both welfare preferences and implementation mistakes from choice data, for a given behavioral model. Several contributions have studied this problem for specific models. Recent examples include Masatlioglu et al. (2012) for a model of limited attention, and Kőszegi and Szeidl (2013) for a model of focusing. See Manzini and Mariotti (2014) for a general overview.
oglu et al. (2012). Even though restrictive, we consider the approach to be a proper starting point for the analysis of nudging, where the regulator selects frames but exercises no control over choice sets. The framework accommodates a large range of behavioral models. Among others, we will study well-known models such as default bias, satisficing, limited search, and choice from lists. Other models that would fit the framework include loss aversion when frames determine the reference point (Tversky and Kahneman, 1991), priming and anchoring (Ariely et al., 2003), dynamic inconsistency when frames capture the time of choice (see Bernheim and Rangel, 2009, p. 57f, for an example), and general behavioral biases in the context of binary choice (Mullainathan et al., 2012). Our goal is not to take a stand on what the correct behavioral model is, or to argue in favor of any one of these models. Instead, the objective of our analysis is to understand the general properties of decision-making processes that make it possible or impossible to improve choices by framing.

A first contribution of our paper is to provide a choice-theoretic definition of a nudge. After identifying the welfare preferences that are consistent with a given data set and a behavioral model, in Section 3 we evaluate the frames on the basis of these preferences. Comparing frames pairwise, we say that frame $f$ is a successful nudge over frame $f'$ if the choices under $f$ are always at least as good as under $f'$ and sometimes strictly better, irrespective of which of the consistent preferences is the actual welfare preference. This definition captures the above-mentioned idea that the regulator aims at improving the agent’s choices by her own standards, i.e., the regulator tries to help the agent do what she really wants to do.

Ideally, we would like to identify a frame that is a successful nudge over all the other frames. We first show that the ability to identify such an optimal frame coincides with the ability to identify the welfare preference in the behavioral model. An optimal frame is revealed by some sufficiently rich data set if and only if the welfare preference is fully revealed by some sufficiently rich data set. This does not mean that the welfare preference has to be fully elicited for successful nudging. We will show that it can indeed happen that the regulator has learned enough to pick an optimal framing but not yet enough to determine the agent’s welfare-optimal choices.

There are interesting examples of behavioral models that have identifiable preferences, such as a limited search model. There are also interesting examples of models with non-identifiable preferences, such as a satisficing model. We can show that identifiable preferences become the generic case when the set of alternatives grows large. Therefore,

---

4See Masatlioglu and Nakajima (2015) for a version of the limited attention model with frames. Other recent examples of behavioral models which generate incoherence across choice sets include Manzini and Mariotti (2007, 2012), Cherepanov et al. (2013) and Lleras et al. (2017). De Clippel and Rozen (2014) investigate the problem of learning from incomplete data sets without imposing the assumption of coherence.
with many alternatives, an optimal nudge can generically be identified even without prior knowledge of the desired behavior. However, we need to qualify this positive interpretation because finding the optimal nudge can still be very difficult. In fact, we show that the informational complexity can become prohibitive.

When an optimal nudge is out of reach, either because welfare preferences are not identifiable or because the complexity of the elicitation problem is too large, then the regulator can still pursue the more modest goal of identifying frames which are dominated by others. Put differently, even when it is impossible to find a frame that improves upon all other frames, it may still be possible that some frames can be improved upon. Such dominated frames can exist, as we show by example. However, if the behavioral model satisfies a property that we term the \textit{frame-cancellation property}, then all frames are always undominated, irrespective of the data set’s richness. Several important models have the frame-cancellation property. A first example is the satisficing model in its different varieties. A second example is the much-discussed case where the agent chooses the one alternative out of two that is marked as a default. We also present a decision-making procedure with limited sensitivity that nests all these (and more) behavioral models. With the frame-cancellation property, observation of choices can reveal a lot about the agent’s welfare preference, but it never reveals the information required to improve these choices. This finding is a striking impossibility result, as it shows that essential information remains inaccessible even when the behavioral model is perfectly known and even when an unlimited amount of behavioral data can be collected. In other words, prior knowledge of the desired behavior is not only sufficient but also necessary for nudging when the frame-cancellation property holds.

Given the impossibility result, in Section 4 we weaken our definition of a successful nudge in two dimensions. First, the previous definition is too stringent if the regulator has additional information about the agent’s choice environment, such as her income and a range of conceivable market prices. Then it is not necessary to require improved choices for all choice sets, but only for those that could be affordable to the agent. Alternatively, the regulator herself may have limited the admissible choice sets, for example because she banned activities like smoking which were ex ante known to cause negative externalities. We show by example that a frame can become an optimal nudge when the environment is restricted adequately. In particular, there can be non-trivial interactions between soft and hard regulatory interventions. Yet, this is never the case when the behavioral model satisfies the frame-cancellation property, because our previous impossibility result is perfectly robust to restricted environments.

The second dimension in which we weaken our definition of a nudge concerns the extent of agreement among the welfare preferences. Our original, cautious definition requires a successful nudge to improve choices for all welfare preferences that are consistent with
the data set. A less cautious regulator may be content to help a sufficiently large share of those preferences. We again show by example how an optimal nudge can sometimes be identified if the unanimity requirement is replaced by a qualified majority requirement. This holds even if the behavioral model satisfies the frame-cancellation property. However, our previous impossibility result comes back under a mild additional property, which requires that the sets of behaviorally equivalent welfare preferences are symmetric in a well-defined sense. Among others, this property is satisfied by the entire class of limited sensitivity models mentioned above. Hence, the difficulties that we have identified for models like satisficing or default bias are not due to excessive caution on the side of the regulator.

Some of our results show that successful nudging can be possible even without prior knowledge of the desired behavior. Other results imply strong limitations for a regulator who attempts to base the selection of nudges on a revealed preference foundation. This is particularly noteworthy given that several of our assumptions work in favor of nudging. We will discuss this again in Section 5, where we relate our results to the literature, in particular the literature on behavioral welfare economics.

Our analysis shows that seemingly minor differences between behavioral models – such as whether an agent’s failure to optimize is due to a low aspiration level as in the satisficing model or due to a restricted number of considered alternatives as in the limited search model – can have profoundly different consequences for the ability to improve well-being by framing. We have to leave open the question whether it will ever be possible to distinguish between models in such detail. At any rate, the analysis points at interesting directions for research on decision-making processes and their normative implications.

2 Model and Examples

We begin by introducing the formal framework, which is a variant of RS, and we illustrate it with the help of two examples.

Let $X$ be a finite set of alternatives, with $m_X = |X|$. Denote by $P$ the set of linear orders (reflexive, complete, transitive, antisymmetric) on $X$. A strict preference is a linear order $\succeq \in P$. We use the symbol $\succeq$ despite the assumed strictness, because preferences are reflexive and we will indeed encounter statements like $x \succeq x$. We use the symbol $\succ$ to denote the asymmetric part of $\succeq$, i.e., we write $x \succ y$ when $x \succeq y$ and $x \neq y$. Let $F$ be a finite set of frames, with $m_F = |F|$. By definition, frames capture all dimensions of the environment that can affect decisions but are not considered welfare-relevant.\(^5\)

\(^5\)For specific applications, the modeller has to judge which dimensions are welfare-relevant and which are not. For instance, it may be uncontroversial that an agent’s well-being with some level of old age savings is independent of whether this level was chosen by default or by opt-in, but analogous statements would not be true if a default entails substantial switching costs, or if a frame actually provides novel
The agent’s behavior is summarized by a distortion function \( d : P \times F \to P \), which assigns a preference \( d(\succeq, f) \in P \) to each combination of \( \succeq \in P \) and \( f \in F \). The interpretation is that an agent with welfare preference \( \succeq \) acts as if maximizing the behavioral preference \( d(\succeq, f) \) if the choice situation is framed by \( f \). As discussed before, this assumes that choices are coherent across choice sets whenever a frame is held fixed. For illustration purposes, we formally introduce two possible models.

**Model 1 (Perfect-Recall Satisficing).** This model is taken from RS. The agent is satisfied with any of the top \( k \) alternatives in her welfare preference, so \( k \in \{2, \ldots, m_X\} \) represents her aspiration level. The frame \( f \) describes the order in which the alternatives are presented to the agent. Whenever the agent chooses from some non-empty choice set \( S \subseteq X \) (e.g. the budget set), she considers the alternatives in \( S \) sequentially in their order as prescribed by \( f \in F = P \). She chooses the first alternative that exceeds her aspiration level, i.e., she picks from \( S \) whichever satisfactory alternative is presented first. If \( S \) turns out not to contain any satisfactory alternative, the agent recalls all alternatives in \( S \) and chooses the welfare-optimal one. Choices between satisfactory alternatives will thus always be in line with the order of presentation, while all other choices are in line with the welfare preference. Hence we can obtain \( d(\succeq, f) \) from \( \succeq \) by rearranging the top \( k \) elements according to their order in \( f \).

**Model 2 (Limited Search).** This model formalizes a choice heuristic similar to one described in Masatlioglu et al. (2012). When the agent looks for a product online, all alternatives in \( X \) are displayed by a search engine, but only \( k \) of them on the first result page and the remaining \( m_X - k \) of them on the second result page. The frame \( f \) here is the set of \( k \in \{1, \ldots, m_X - 1\} \) alternatives on the first page, such that \( F \) is the set of all size \( k \) subsets of \( X \). The agent again chooses from non-empty choice sets \( S \subseteq X \) (e.g. not all displayed alternatives may be affordable to the agent or in stock with the retailer). Whenever the first result page contains at least one of the alternatives from \( S \), then the agent does not look at the second page but chooses from \( S \cap f \) according to her welfare preference. Only if none of the elements of \( S \) is displayed on the first page, then the agent moves to the second page and chooses there according to her welfare preference. Choices between alternatives on the same page will thus always be in line with the welfare preference, but any available alternative on the first page is chosen over any alternative on the second page. Hence \( d(\succeq, f) \) preserves \( \succeq \) among all first and among all second page alternatives, but takes the first page to the top.

---

6In contrast to RS, we explicitly treat the order of presentation as a variable frame. We also assume that the aspiration level \( k \) is fixed, which implies that the distortion function is single-valued.
The function $d$ should be thought of as the regulator’s conjecture about the relation between welfare, frames and choice. Such a conjecture will typically rely on insights about the decision-making process and thus originates from non-choice data.\(^7\) For instance, eye-tracking or the monitoring of browsing behaviors could provide the type of information necessary to substantiate a model like limited search (see the discussion in Masatlioglu et al., 2012). Methods from neuroscience may confirm decision-processes such as perfect-recall satisficing. Furthermore, model conjectures can be falsified with the help of choice data (see RS and Manzini and Mariotti, 2014).

As noted before, it is not our goal here to argue that a specific behavioral model $d$ is correct. Rather, the objective of our analysis is to understand the general properties of decision-making processes that make it possible or impossible to improve choices by framing. Hence the only minor assumption that we impose on the behavioral model is that for each $\succeq \in P$ there exists an $f \in F$ such that $d(\succeq, f) = \succeq$. This rules out that some welfare preferences are distorted by all possible frames, and therefore the scope for nudging is not exogenously constrained. The assumption does not imply the existence of a neutral frame that is non-distorting for all preferences.\(^8\) In the satisficing model, all frames which present the $k$ satisfying alternatives in their actual welfare order are non-distorting for that welfare preference. In the limited search model, the non-distorting frame places the $k$ welfare-best alternatives on the first page. Note that the framing indeed has to be non-distorting if we want to guarantee welfare optimal choices from all choice sets $S \subseteq X$, while even a distorting frame may induce the optimal choice from specific choice sets.

Holding fixed a frame, the regulator now observes the agent’s choices from sufficiently many choice sets $S \subseteq X$ to deduce her behavioral preference, in the usual revealed preference sense. Here the only difference to the usual approach is that the behavioral preference is not automatically equated with the welfare preference, and that the procedure generates potentially different behavioral preferences when repeated for different frames.

Formally, a data set is a subset $\Lambda \subseteq P \times F$, where $(\succeq', f') \in \Lambda$ means that the agent has been observed under frame $f'$ and her choice behavior revealed the behavioral preference $\succeq'$. Further following RS, we say that $\succeq$ is consistent with data set $\Lambda$ if for each $(\succeq', f') \in \Lambda$ it holds that $\succeq' = d(\succeq, f')$. In that case, $\succeq$ is a possible welfare preference.

---

\(^7\)Arguably, non-choice-based conjectures about the relation between choice and welfare always have to be invoked, even in standard welfare economics, see Köszegi and Rabin (2007, 2008a), Rubinstein and Salant (2008) and Manzini and Mariotti (2014). For an opposing perspective and a critical discussion of the ability to identify the decision process, see Bernheim (2009).

\(^8\)Sometimes a neutral or “revelatory” frame (Goldin, 2015, p. 9) may indeed exist, for example when the default can be removed from a choice problem. If such a frame exists and is thus known to the regulator (due to the assumption that the distortion function $d$ is known), the solution to the welfare elicitation problem and also the nudging problem becomes straightforward. Often, however, this solution is not available, e.g. defaults are unavoidable for organ donations, alternatives must always be presented in some order or arrangement, and questions must be phrased in one way or another.
because the data set could have been generated by an agent with that preference. We now illustrate this elicitation of the welfare preference, and also some first implications for nudging, using our two examples.

Example 1. Consider an agent whose decision process is described by the perfect-recall satisficing model with aspiration level $k = 2$. The set of alternatives is given by $X = \{a, b, c, d\}$. The agent has the welfare preference $\succeq_1$ given by $c \succ_1 a \succ_1 b \succ_1 d$, so that the alternatives $c$ and $a$ are satisfactory. Denote the frame which presents the alternatives in alphabetical order by $f$. Thus, when choosing from some choice set $S \subseteq X$, the agent will consider the alternatives in $S$ in alphabetical order and choose the first which is satisfactory. Consequently, because $a$ is presented before $c$, the agent will choose $a$ whenever $a \in S$, even if also $c \in S$, in which case this is a mistake. She will choose $c$ when $c \in S$ but $a \notin S$, and otherwise she will choose $b$ over $d$ by the perfect-recall assumption. Taken together, these choices look as if the agent was maximizing the preference $\succeq_2$ given by $a \succ_2 c \succ_2 b \succ_2 d$. Formally, we have $d(\succeq_1, f) = \succeq_2$.

Suppose the behavioral preference $\succeq_2$ is observed in the standard revealed preference sense, by observing the agent’s choices from different choice sets $S \subseteq X$ but under the fixed frame of alphabetical presentation. Formally, the regulator obtains the data set $\Lambda = \{(\succeq_2, f)\}$. Given the perfect-recall satisficing conjecture, she can then conclude that the agent’s welfare preference must be either $c \succ_1 a \succ_1 b \succ_1 d$ or $a \succ_2 c \succ_2 b \succ_2 d$; these two but no other welfare preferences generate the observed behavior under frame $f$. Formally, the set of preferences that are consistent with the data set is given by $\{\succeq_1, \succeq_2\}$. Therefore, with as little information as observing behavior under a single frame, the set of possible welfare preferences can be reduced from initially 24 to only 2.

We now illustrate some first implications for nudging, which here amounts to fixing an optimal order of presentation. Any order that presents $a$ before $c$ would be optimal if the agent’s welfare preference was $a \succ_2 c \succ_2 b \succ_2 d$ but induces the above described decision mistake between $a$ and $c$ if the welfare preference is $c \succ_1 a \succ_1 b \succ_1 d$. The exact opposite is true for any order that presents $c$ before $a$. Hence our knowledge is not yet enough to favor any one frame over another. As it turns out, the problem cannot be solved by observing the agent under additional frames. The order of presentation fully determines choices among the alternatives $a$ and $c$, so we can never learn about the welfare preference between the two. Since precisely this knowledge would be necessary to determine the optimal order, nudging here runs into irresolvable information problems.

This framework corresponds to the extension in RS where behavioral data sets contain information about frames. It simplifies their setup by assuming that any pair of a welfare preference and a frame generates a unique distorted behavioral preference. This is not overly restrictive, as the different contingencies that generate a multiplicity of distorted preferences can always be written as different frames. It is restrictive in the sense that observability and controllability of these frames might not always be given.
Example 2. Consider an agent whose decision process is described by the limited search model, and \( k = 2 \) alternatives are presented on the first result page. As in the previous example, the set of alternatives is \( X = \{a, b, c, d\} \) and the agent has the welfare preference \( \succeq_1 \) given by \( c \succ_1 a \succ_1 b \succ_1 d \). Let \( f = \{a, b\} \) denote the frame which puts the alternatives \( a \) and \( b \) on the first page. Thus, whenever the agent’s choice set \( S \subseteq X \) contains either \( a \) or \( b \) (or both), she will remain on the first page and make her choice there. Consequently, she chooses \( a \) whenever \( a \in S \), even if also \( c \in S \), because \( c \) is displayed only on the second page. This is again a mistake. She will choose \( b \) when \( b \in S \) but \( a \notin S \), and otherwise she will choose \( c \) over \( d \). Taken together, these choices look as if the agent was maximizing the preference \( \succeq_3 \) given by \( a \succ_3 b \succ_3 c \succ_3 d \). Formally, we have \( d(\succeq_1, f) = \succeq_3 \).

Suppose again that this behavioral preference is revealed, i.e., the regulator obtains the data set \( \Lambda = \{(\succeq_3, f)\} \). Reversing the distortion process now unveils that the agent truly prefers \( a \) over \( b \) and \( c \) over \( d \), which leaves the six possible welfare preferences marked in the first column of Table 1. The set of preferences consistent with the observed behavior is therefore given by \( \{\succeq_1, \succeq_2, \succeq_3, \succeq_4, \succeq_5, \succeq_6\} \), meaning that the single observation reduces the set of possible welfare preferences from 24 to 6.

<table>
<thead>
<tr>
<th>( c \succ a \succ b \succ d )</th>
<th>( f = {a, b} ): ( a \succ_3 b \succ_3 c \succ_3 d )</th>
<th>( f' = {a, d} ): ( a \succ_7 c \succ b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a \succ c \succ b \succ d )</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( a \succ c \succ d \succ b )</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( c \succ a \succ b \succ d )</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( c \succ a \succ d \succ b )</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( c \succ d \succ a \succ b )</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( a \succ c \succ a \succ b )</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: Reversing Limited Search

Here, an optimal nudge should place the two welfare-best alternatives on the first page, thus helping the agent avoid decision mistakes like the one between \( a \) and \( c \) under frame \( f \). Unfortunately, each of the four alternatives still belongs to the top two for at least one of the consistent welfare preferences, but none of them for all of the consistent welfare preferences. Hence no frame guarantees fewer mistakes than any other.

In contrast to the satisficing example, however, gathering more information helps. Observing choices under frame \( f' = \{a, d\} \) reveals the behavioral preference \( \succeq_7 \) given by \( a \succ_7 c \succ_7 b \), from which the six welfare candidates marked in the second column of Table 1 can be deduced. Formally, adding this observation to the data set yields \( \Lambda' = \{(\succeq_3, f), (\succeq_7, f')\} \), and the set of consistent welfare preferences shrinks to \( \{\succeq_1, \succeq_2, \succeq_4, \succeq_5\} \). Note that these preferences all agree that \( a \) and \( c \) are the two best alternatives. Hence we know that \( f'' = \{a, c\} \) is the optimal nudge.
To find the optimal nudge, it was sufficient to observe behavior under only two frames. The actual welfare preference is still not known, so the example also shows that identifying a nudge is not the same problem as identifying the welfare preference. Based on data set $\Lambda'$, the regulator does not know the agent’s optimal alternative in each choice set, and therefore she would be unable to choose in the interest of the agent. However, she knows enough to determine the optimal framing and delegate the choice to the agent. This is a fundamental informational argument for soft rather than hard paternalism.

We conclude this section with some remarks on our framework. Our strong assumption of a given, unique conjecture $d$ presumes substantial knowledge of the regulator regarding the distortions that frames induce. In practice, a regulator may know only some aspects of how frames affect behavior. Further, we implicitly assume that the regulator can perfectly observe and control the frames. Some dimension of frames may reflect internal states of the agent, such as a person’s mood, and these are typically not fully under the regulator’s control. Finally, it is rare to observe choices from sufficiently many different subsets of alternatives to fully deduce the behavioral preference of an agent. While this assumption is also implicit in the standard revealed preference approach, it is not an ability that a regulator generally has. For the interpretation of the following results, it should be kept in mind that these assumptions tend to help the regulator and stack the odds in favor of nudging.\footnote{Having said this, it is of course possible to relax some of these assumptions. See Appendix B for a formal analysis of uncertainty about the model $d$ and imperfectly observable frames.}

### 3 Main Results

#### 3.1 Definition of a Nudge

In this section, we will provide a formal definition of a nudge. To capture the first step of preference elicitation due to RS in a concise way, let

\[
\bar{\Lambda}(\succeq) = \{(d(\succeq, f), f) \mid f \in F\}
\]

be the maximal data set that could be observed if the agent’s welfare preference was $\succeq$, i.e., the data set that contains an observation for each possible frame. Then the set of all welfare preferences that are consistent with an arbitrary data set $\Lambda$ can be written as

\[
P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\}.
\]

Without further mention, we consider only data sets $\Lambda$ for which $P(\Lambda)$ is non-empty, i.e., for which there exists $\succeq$ such that $\Lambda \subseteq \bar{\Lambda}(\succeq)$. Otherwise, the behavioral model would be...
falsified by the data. Observe that a frame $f$ cannot appear more than once in such data sets. Observe also that $P(\emptyset) = P$ holds, and that $P(\Lambda) \subseteq P(\Lambda')$ whenever $\Lambda' \subseteq \Lambda$.

We are interested in evaluating the frames after having observed some data set $\Lambda$ and having narrowed down the set of possible welfare preferences to $P(\Lambda)$. Since previously different frames may now have become behaviorally equivalent, let

$$[f]_\Lambda = \{ f' \mid d(\succeq, f') = d(\succeq, f), \forall \succeq \in P(\Lambda) \}$$

be the equivalence class of frame $f$, so the elements of $[f]_\Lambda$ induce the same behavior as $f$ for all of the remaining possible welfare preferences. We denote by $F(\Lambda) = \{ [f]_\Lambda \mid f \in F \}$ the quotient set of all equivalence classes. Our central definition compares the elements of $F(\Lambda)$ pairwise from the perspective of the possible welfare preferences. For any $\succeq$ and any non-empty $S \subseteq X$, let $c(\succeq, S)$ be the element of $S$ that would be chosen from $S$ by an agent who maximizes $\succeq$.

**Definition 1** For any data set $\Lambda$ and any two frames $f$ and $f'$ such that $[f]_\Lambda \neq [f']_\Lambda$, frame $f$ is a successful nudge over frame $f'$, written

$$[f]_\Lambda N(\Lambda) [f']_\Lambda,$$

if for each $\succeq \in P(\Lambda)$ it holds that $c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)$, for all non-empty $S \subseteq X$.

The statement $[f]_\Lambda N(\Lambda) [f']_\Lambda$ means that the agent’s choice under frame $f$ (and all equivalent ones) is at least as good as under frame $f'$ (and all equivalent ones), in all choice sets and no matter which of the remaining welfare preferences is the true one. Furthermore, there is at least one choice set and one welfare preference for which the choice is strictly improved (this follows because $f$ and $f'$ are not behaviorally equivalent). Welfare preferences enter this definition not only for the evaluation of choices, but also because agents with different preferences may react differently to the same frame.

The binary nudging relation $N(\Lambda)$ shares with other approaches in behavioral welfare economics the property of requiring agreement among multiple preferences (see, for instance, the multiself Pareto interpretation of the unambiguous choice relation by Bernheim and Rangel, 2009). The multiplicity of welfare preferences here simply reflects a lack of better information. As in Masatlioglu et al. (2012), our definition embodies a cautious approach and ensures that the regulator does not accidentally make the agent worse off. We will investigate relaxations of this approach in Section 4 below.\footnote{As a very different revealed preference approach to the problem of ranking frames, one may contemplate the idea of letting the agent choose between the frames. However, this would only shift the problem to a higher level, because the choice between frames would have to be framed in one way or another.}

Technically, the binary relation $N(\Lambda)$ on $F(\Lambda)$ is asymmetric and transitive, hence a strict partial order. It will typically not be complete. But note that adding observations
to a data set can only make the partition $F(\Lambda)$ coarser and the nudging relation more complete, because it can only reduce the set of possible welfare preferences for which improved choices have to be guaranteed. In fact, the only way in which the data set matters for the nudging relation is via the induced set $P(\Lambda)$.

### 3.2 Optimal Nudge

A benevolent regulator would ideally like to choose a frame that guarantees the best possible choices, which means that it should be a successful nudge over all other frames. We call such a frame an optimal nudge. Given a data set $\Lambda$, let

$$G(\Lambda) = \{ f \mid [f]_\Lambda N(\Lambda)[f']_\Lambda \text{ for all } [f']_\Lambda \neq [f]_\Lambda \}$$

be the set of frames which have been identified as optimal. Formally, $G(\Lambda)$ coincides with the greatest element of the strict partially ordered set $F(\Lambda)$, and it might be empty due to incompleteness of the binary nudging relation. Since the nudging relation becomes more complete as we collect additional observations, it follows that optimal nudges are more likely to exist for larger data sets. Therefore, we now provide a necessary and sufficient condition for the existence of an optimal nudge based on maximal data sets.

**Definition 2** Welfare preference $\succeq$ is identifiable if for each $\succeq' \in P$ with $\succeq' \neq \succeq$, there exists $f \in F$ such that $d(\succeq, f) \neq d(\succeq', f)$.

**Proposition 1** $G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is identifiable.

The if statement is immediate: an identifiable welfare preference is known perfectly once the maximal data set has been collected, and all the non-distorting frames are optimal with that knowledge. The only-if statement tells us that there is no hope to ever identify an optimal nudge if the welfare preference cannot be identified, i.e., if there exists another welfare preference $\succeq'$ that is behaviorally equivalent to $\succeq$ under all the frames. In that case we say that $\succeq$ and $\succeq'$ are indistinguishable. A frame can of course be optimal even if two (or more) possible welfare preferences are left, but it cannot be simultaneously optimal for two indistinguishable welfare preferences, because behavior under the optimal frame would allow to empirically discriminate between the two preferences.

It is worth emphasizing again that the result does not imply that the welfare preference actually has to be learned perfectly to find an optimal nudge. It only tells us that, if $\succeq$ is the true and identifiable welfare preference, then for some sufficiently large data set $\Lambda$ we will be able to identify an optimal nudge. The set $P(\Lambda)$ can still contain more than one element at that point. This observation is important, as the case for soft versus hard paternalism is particularly compelling when there is a reason for giving freedom of choice.
to the agent. Such a reason exists whenever the regulator knows enough to determine the optimal nudge but not yet the agent’s optimal choices, as in our Example 2.

Proposition 1 shows that prior knowledge of the desired behavior is not necessary for nudging when the behavioral model has identifiable preferences, by which we mean from now on that all welfare preferences are identifiable. The relevant information for nudging can then always be elicited from the agent’s choice behavior. Conversely, prior knowledge is necessary when the behavioral model does not have identifiable preferences.

A natural question that arises from these findings is about the relevance of models with or without identifiable preferences. Our prime example for non-identifiable preferences is the perfect-recall satisficing model. Any two welfare preferences that are identical except that they rank the same best \( k \) alternatives differently are mapped into the same distorted preference by any frame, and hence are indistinguishable. Our prime example for identifiable preferences is the limited search model (for \( m_X \geq 3 \)). We learn the welfare preference among all alternatives on the same result page, and thus we can identify the complete welfare preference by observing behavior under sufficiently many different frames. The decision processes formalized by these two models both appear plausible, suggesting that both classes of models may be important.

A more quantitative approach is to ask about the genericity of the two classes. For any fixed numbers \( m_X \) of alternatives and \( m_F \) of frames, there is a finite number of possible models \( d \). Suppose we draw one model randomly from among all these models, with equal probabilities. How likely will it have identifiable or non-identifiable preferences? Below, we provide an answer to this question for the limiting case as the number of alternatives grows large.\(^{12}\) For the analysis, we need to specify how the number of frames varies with the number of alternatives. The two numbers should typically be increasing together. For instance, if we restricted attention to models where frames are orders of presentation, we would already obtain \( m_F = m_X! \).\(^{13}\) For the following result, however, we only need to assume that \( m_F \geq 4 \) for sufficiently large values of \( m_X \).

Proposition 2 The share of models with identifiable preferences goes to 1 as \( m_X \to \infty \).

It is difficult to determine precisely the share of models with identifiable preferences among all possible models, but in the proof we find a lower bound that is tractable and converges to 1 as the number of alternatives grows. Roughly speaking, with more alternatives and thus more preferences, there are more ways in which each welfare preference can distinguish itself behaviorally from the others.

---

\(^{12}\)Our approach here is similar to Kalai et al. (2002), who are interested in the number of different preferences that are necessary to rationalize a randomly drawn choice function. They show that the share of all conceivable choice functions which can be rationalized by less than some maximal number goes to 0 as the number of alternatives grows large.

\(^{13}\)In general, the number of frames can be arbitrarily large, but there can never be more than \( m_X!m_X! \) different non-equivalent frames, the number of mappings from \( P \) to \( P \).
If one accepts that model genericity captures model relevance in a meaningful way, then the result is good news for the nudging project. When the number of alternatives is large, an optimal nudge can generically be identified.

We note, however, that the complexity of finding the optimal nudge may still become prohibitive. Observing a single behavioral preference already becomes more demanding with a larger number of alternatives, a problem that also arises in the standard revealed preference approach. This problem is magnified here, because behavioral preferences have to be observed under multiple frames. We investigate this particular problem formally in Appendix B. We assume there that a regulator can impose a sequence of frames on the agent during an observation stage, with the goal of eliciting the necessary information for nudging efficiently. We define the complexity \( n \) of the problem as the number of frames that the quickest data gathering procedure must impose until it identifies an optimal nudge for sure. This number can be surprisingly small for specific behavioral models. For instance, we show that \( n \leq 3 \) for the limited search model, so this model’s complexity does not even increase in the number of alternatives. However, we also establish a tight bound on \( n \) for arbitrary behavioral models. The bound grows more than exponentially in the number of alternatives, which implies that the informational requirements can become prohibitive even with identifiable welfare preferences.

### 3.3 Dominated Frames

When an optimal nudge is out of reach, either because welfare preferences are not identifiable or because the informational complexity of the problem is too large, we might still be able to exclude some frames that are dominated by others. Formally, let

\[
D(\Lambda) = \{ f \mid [f']_\Lambda N(\Lambda)[f]_\Lambda \text{ for some } [f']_\Lambda \neq [f]_\Lambda \}
\]

be the set of frames which have been identified as being dominated. \( D(\Lambda) \) is the union of all elements that are not maximal in the strict partially ordered set \( F(\Lambda) \). To provide an analogy, we can think of these frames as Pareto inefficient policies. A frame \( f \) that is dominated can be safely excluded, as there exists a nudge \( f' \) which ensures that no possible type of agent is made worse off, while at least one is made strictly better off.

Dominated frames can exist already ex ante, with no data on the agent’s behavior.\(^{14}\) For instance, certain informational arrangements could be interpreted as being dominant over others, because they objectively clarify the available information and can only im-

\(^{14}\)Formally, this means that \( D(\emptyset) \) can be non-empty. As a simple example, suppose frame \( f' \) makes a particular alternative \( x \) salient and moves it up by exactly one position in the behavioral preference relative to the welfare preference (if still possible). Suppose that the effect of frame \( f \) is even stronger, so that \( x \) is moved up by two positions. Then frame \( f' \) induces weakly better choices than \( f \) for all welfare preferences, and strictly better choices for some welfare preferences, which implies \( f \in D(\emptyset) \).
prove the decision quality (e.g. Camerer et al., 2003). In the following example, we show that ex ante undominated frames can become dominated with richer data, too. For reasons that will become apparent later, the example is abstract rather than describing a plausible decision process like satisficing.

**Example 3.** Assume that $X = \{a, b, c, d\}$ and $F = \{f_1, f_2, f_3\}$. Consider the distortion function among the five preferences depicted in Figure 1. The simplifying assumption that all other 19 welfare preferences are left completely undistorted by each of the three frames allows us to focus exclusively on these five preferences. The two welfare preferences $\succeq_1$ and $\succeq_2$ are indistinguishable, as each frame maps them into the same behavioral preference. The same holds for $\succeq_3, \succeq_4$ and $\succeq_5$, which are also mutually indistinguishable.

Note first that no frame is dominated before any data has been collected, because each one is the unique non-distorting frame for one possible welfare preference. Now suppose we observe $\Lambda = \{(\succeq_1, f_1)\}$ so that $P(\Lambda) = \{\succeq_1, \succeq_2\}$. It follows immediately that none of the potentially non-distorting frames $f_1$ and $f_2$ can be dominated. The frame $f_3$, however, is now dominated by $f_1$. If the welfare preference is $\succeq_2$, then $f_1$ induces a decision mistake between $a$ and $b$, but so does $f_3$, which induces an additional decision mistake between $c$ and $d$. Hence we obtain $D(\Lambda) = \{f_3\}$. We have learned enough to identify a nudge over $f_3$, but no additional observation will ever allow us to rank $f_1$ and $f_2$. We will use the additional structure in the example later, to illustrate some results in Section 4.

![Figure 1: Abstract Distortion Function.](image)

The sometimes dominated frame $f_3$ in Example 3 has a particular property. It maps the indistinguishable set of preferences $\{\succeq_1, \succeq_2\}$ outside of itself. This is the reason why the example violates the following property.

**Definition 3** A distortion function $d$ has the frame-cancellation property if

$$d(d(\succeq, f_1), f_2) = d(\succeq, f_2)$$

holds for all $\succeq \in P$ and all $f_1, f_2 \in F$. 
With the frame-cancellation property, the impact of any frame \( f_1 \) disappears once a new frame \( f_2 \) is applied. Starting from any welfare preference \( \succeq \), the preference \( d(\succeq, f) \) obtained by applying any frame \( f \) is then itself a welfare preference that is observationally equivalent to \( \succeq \). As a consequence, all maximal indistinguishable sets of preferences are closed under the distortion function for any frame, in contrast to Example 3.

Many interesting behavioral models have the frame-cancellation property. An extreme example, where frames never have an effect on behavior and \( d(\succeq, f) = \succeq \) always holds, is the rational choice model. All welfare preferences are identifiable in the rational choice model, which constitutes of course the basis for the standard revealed preference approach. We note that the rational choice model is the only model which has both identifiable preferences and the frame-cancellation property.\(^{15}\) The opposite extreme case of frame-cancellation arises when \( d(\succeq, f) \) is independent of \( \succeq \), so that frames override the preference entirely. This is true, for instance, when there are only two alternatives and the agent always chooses the one that is marked as the default.

The perfect-recall satisficing model has the frame-cancellation property, too, even though the welfare preference retains a substantial impact on behavior. In this model, the effect of the order of presentation is to overwrite the welfare preference among the top \( k \) alternatives, which leaves no trace of previous frames when done successively.

We can also establish a connection to the analysis of choice from lists by Rubinstein and Salant (2006). They allow for the possibility that agents choose from lists instead of sets, i.e., the choice from a given set of alternatives can be different when the alternatives are listed differently. Their results imply that we can capture choice from list behavior in reduced form of a distortion function whenever the axiom of “partition independence” is satisfied by the agent’s choices for all possible welfare preferences.\(^{16}\) An example in which this holds is satisficing without recall. In contrast to the perfect-recall version, the agent here chooses the last alternative on a list when no alternative on the list exceeds her aspiration level. Formally, \( d(\succeq, f) \) is obtained from \( \succeq \) by rearranging the top \( k \) elements in the order of \( f \) and the bottom \( m_X - k \) elements in the opposite order of \( f \) (see RS). It is easy to verify that this model has the frame-cancellation property.

The following general class of decision processes nests all these models with the frame-cancellation property.

\(^{15}\)To see why, suppose \( d \) is not fully rational, i.e., there exist \( \succeq' \) and \( f' \) such that \( d(\succeq', f') \neq \succeq'' \). If \( d \) has the frame-cancellation property, we then obtain \( d(\succeq'', f) = d(\succeq', f') \) for all \( f \in F \), hence \( \succeq' \) and \( \succeq'' \) are indistinguishable and not identifiable. It is possible to have models with the frame-cancellation property that are semi-rational, in the sense that some but not all welfare preferences are unaffected by frames. In that case, exactly those welfare preferences which always act rationally are identifiable, while the others are not.

\(^{16}\)Partition independence requires that the choice from two concatenated sublists is the same as the choice from the list that concatenates the two elements chosen from the sublists (Rubinstein and Salant, 2006, p. 7). Such behavior can be modelled as the maximization of some non-strict preference that is turned strict by ordering its indifference sets in or against the list order (Proposition 2, p. 8).
Model 3 (Limited Sensitivity). The agent displays limited sensitivity in the sense that she can sometimes not tell whether an alternative is actually better than another. Degree and allocation of sensitivity are described by a vector $\kappa = (k_1, k_2, \ldots, k_t)$ of positive integers with $\sum_{i=1}^{t} k_i = m_X$. A welfare preference $\succeq$ induces an ordered partition of $X$, where block $X_1$ contains the $k_1$ welfare-best alternatives, $X_2$ contains the $k_2$ next best alternatives, and so on. The agent can distinguish alternatives across but not within blocks. When choosing from $S \subseteq X$, she therefore only identifies the smallest $i$ for which $S \cap X_i$ is non-empty, and the frame then fully determines the choice from this set. Thus $d(\succeq, f)$ is obtained from $\succeq$ by rearranging the alternatives within each block of the partition in a way that does not depend on their actual welfare ranking. Formally, for any $\succeq$ let $P_{\succeq}^\kappa$ be the set of welfare preferences that induce the same ordered partition of $X$ as $\succeq$. Then $d(\succeq', f) = d(\succeq'', f) \in P_{\succeq}^\kappa$ must hold whenever $\succeq', \succeq'' \in P_{\succeq}^\kappa$, for all $f \in F$. Any such function satisfies the frame-cancellation property.\(^{17}\) When $f$ is an order of presentation and the alternatives within each block of the partition are rearranged in or against this order – because the agent chooses the first or the last among seemingly equivalent alternatives – then the process is a successive choice from list model (see Rubinstein and Salant, 2006, for a definition). Special cases include rational choice for the vector $(k_1, k_2, \ldots, k_t) = (1, 1, \ldots, 1)$, perfect-recall satisficing for $(k, 1, \ldots, 1)$, no-recall satisficing for $(k, m_X - k)$, and situations where the welfare preference has no impact on behavior for $k_1 = m_X$.

Given the range of interesting examples of models with the frame-cancellation property, the following result is particularly striking.

Proposition 3 If $d$ has the frame-cancellation property, then the binary nudging relation $N(\Lambda)$ is empty for all $\Lambda$.

With the frame-cancellation property, whenever we consider two frames that are not behaviorally equivalent, then we cannot rank them. In particular, no frame is ever dominated. Irrespective of how many data points we have collected, we will never know enough to improve upon any given frame. Intuitively, since preference $d(\succeq, f)$ is behaviorally indistinguishable from $\succeq$ when the frame-cancellation property holds, and $f$ is non-distorting for $d(\succeq, f)$, each frame $f$ is always optimal for at least one of the remaining possible welfare preferences. It follows a fortiori that there cannot be an optimal nudge whenever frames indeed affect behavior. A special case arises only when all frames are

\(^{17}\)For any $\succeq \in P$, since $\succeq \in P_{\succeq}^\kappa$ holds we have $d(\succeq, f_1) \in P_{\succeq}^\kappa$ for any $f_1 \in F$. Then we also obtain $d(d(\succeq, f_1), f_2) = d(\succeq, f_2)$ for any $f_2 \in F$, which is the frame-cancellation property. We note that there are models with the frame-cancellation property that do not belong to the class of limited sensitivity models. Any model with frame-cancellation allows us to partition $P$ into maximal indistinguishable sets of preferences very similar to the sets $P_{\succeq}^\kappa$ in the limited sensitivity model, but these sets will not in general be generated by some vector $\kappa = (k_1, k_2, \ldots, k_t)$ as required by the limited sensitivity model.
behaviorally equivalent, because all frames are then equally optimal nudges. This can, however, only happen if choice is rational.\footnote{To see why, assume that \([f]_\Lambda = F\) and consider any \(\succeq \in P(\Lambda)\). Since there exists a frame \(f'\) for which \(d(\succeq, f') = \succeq\) holds, the behavioral equivalence of all frames implies that \(d(\succeq, f) = \succeq\) for all \(f \in F\). Hence behavior is rational for all remaining welfare preferences. In the standard rational choice model, this is true already before any data has been collected. In the semi-rational models discussed in footnote 15, it becomes true as soon as one of the rationally acting welfare preferences has been identified.}

\section{Extensions}

\subsection{Weaker Definition of a Nudge}

Our previous definition required that a nudge must improve choices in all choice sets and for all remaining welfare preferences. While normatively desirable, this is also a strong requirement. We now show how it can be weakened, and what this implies for models for which our analysis revealed fundamental informational obstacles, e.g. models with the frame-cancellation property.

Suppose the regulator strives to improve choices only for some choice situations, not necessarily for all of them. For instance, she may have information about the agent’s income and a range of possible market prices, and therefore she insists on improved choices only among the alternatives that could be contained in the agent’s budget set. Alternatively, the regulator herself may have banned some alternatives. Picking up the discussion in the introduction, she may have used prior information to prohibit the choice of alternatives that are known to cause strong externalities and decrease public welfare. A question arising in that case is whether the conventional hard policy interacts with soft paternalism in an interesting way, i.e., whether coercion can be accompanied by nudging to improve welfare further. Suppose in addition that the regulator is willing to select one frame over another already when it improves choices for a sufficiently large share of the welfare preferences. Our previous unanimity rule corresponds to one extreme case of this approach. It could be weakened towards the other extreme of a simple majority rule.

Formally, let \(E\) be a non-empty collection of non-empty choice sets \(S \subseteq X\), describing the environment in which the regulator wants to improve choices. We can then redefine behavioral equivalence classes of frames as

\[ [f]_\Lambda = \{ f' \mid c(d(\succeq, f'), S) = c(d(\succeq, f), S), \forall \succeq \in P(\Lambda) \text{ and } \forall S \in E \}, \]

where dependence on the environment \(E\) is suppressed for notational simplicity. Given any two frames \(f\) and \(f'\), each welfare preference \(\succeq \in P(\Lambda)\) now falls into one of four categories. The first category contains all \(\succeq\) for which \(f\) induces better choices than \(f'\),
which means that
\[ c(d(\succeq, f), S) \geq c(d(\succeq, f'), S) \text{ for all } S \in E, \]
where the improvement is strict for at least one choice set \( S \in E \). The second category contains all welfare preferences for which \( f' \) likewise induces better choices than \( f \). The third category contains the preferences for which \( f \) and \( f' \) cannot be compared, because each of them sometimes induces strictly better and sometimes strictly worse choices. The fourth and final category contains the preferences for which \( f \) and \( f' \) are behaviorally equivalent.

The total number of preferences in the first three categories is strictly positive if and only if \( [f]_\Lambda \neq [f']_\Lambda \), because the choice of \( f \) versus \( f' \) then matters for at least one possible welfare preference. In that case, we denote by \( r(\Lambda)(f, f') \) the number of preferences in the first category divided by the total number of preferences in the first three categories. Thus, \( r(\Lambda)(f, f') \in [0,1] \) is the share of preferences for which \( f \) strictly outperforms \( f' \), among those preferences for which the selection of \( f \) versus \( f' \) actually matters.

**Definition 4** For any data set \( \Lambda \) and any two frames \( f \) and \( f' \) such that \( [f]_\Lambda \neq [f']_\Lambda \), frame \( f \) is a \( q \)-successful nudge in \( E \) over frame \( f' \), written
\[ [f]_\Lambda N_{q,E}(\Lambda) [f']_\Lambda, \]
if \( r(\Lambda)(f, f') \geq q \), where \( 1/2 < q \leq 1 \).

For \( f \) to be a \( q \)-successful nudge over \( f' \) in environment \( E \), it must be the case that \( f \) improves choice from the sets \( S \in E \) for at least a share \( q \) of the preferences that are responsive to the change in framing. The previous unanimity requirement is a special case of this formulation for \( q = 1 \). On the other hand, for values of \( q \) that are sufficiently close to 1/2, we obtain the weak requirement that only a strict majority of welfare preferences must be induced to make better choices. The fact that \( q > 1/2 \) ensures that the relation \( N_{q,E}(\Lambda) \) is still asymmetric and captures strict improvements in a reasonable way.\(^{19}\)

### 4.2 Nudging in Restricted Environments

Let us first maintain the unanimity requirement \( q = 1 \) but work with a restricted set \( E \) of choice sets. As mentioned above, a particularly interesting application is one where

\(^{19}\)We note that \( N_{q,E}(\Lambda) \) is not guaranteed to be transitive when \( q < 1 \), because very different subsets of the welfare preferences can be affected as we move along a sequence of frames. This observation just mirrors the well-known fact that pairwise voting procedures can generate Condorcet cycles when they require less than unanimous assent. It is also no longer true that \( N_{q,E}(\Lambda) \) becomes more complete as \( \Lambda \) grows, because a frame \( f \) can cease to be a \( q \)-successful nudge over \( f' \) once some welfare preferences for which \( f \) outperforms \( f' \) have been ruled out by additional data.
these restrictions are imposed by the regulator for reasons outside of our model. As the following examples illustrates, there can be non-trivial interactions between such hard interventions and nudging.

**Example 3 continued.** Reconsider the behavioral model depicted in Figure 1. Suppose the regulator has banned alternatives $a$ and $b$, so that the agent’s only relevant choice is between $c$ and $d$. This can be formalized as environment $E = \{\{c, d\}\}$. The welfare preferences $\succeq_1$ and $\succeq_2$ agree that $c$ is better than $d$, while $\succeq_3$, $\succeq_4$ and $\succeq_5$ agree that $d$ is better than $c$. Hence, ex ante the regulator still does not know the agent’s optimal alternative. Even before any data has been collected, however, choice under frame $f_3$ is now dominated by choice under either frame $f_1$ or $f_2$. The latter two frames, which are behaviorally equivalent in $E$ for all welfare preferences, induce a better choice than $f_3$ for $\succeq_1$ and $\succeq_2$, and they induce the same choice as $f_3$ for $\succeq_3$, $\succeq_4$ and $\succeq_5$. This is true although $f_3$ is non-distorting for $\succeq_5$, but the choice sets on which $f_3$ would outperform the other frames have been ruled out. Consequently, the coercive policy of banning $a$ and $b$ should be accompanied by a nudge towards $f_1$ or $f_2$.

As the following result shows, it is not possible to make similar recommendations when the behavioral model satisfies the frame-cancellation property.

**Proposition 4** If $d$ has the frame-cancellation property, then the binary nudging relation $N_{1,E}(\Lambda)$ is empty for all $E$ and $\Lambda$.

The proposition generalizes our previous result and establishes its perfect robustness for restricted environments. Our earlier intuition is still valid. It remains impossible to identify any dominated frame because framing effects conceal exactly the information that we would need to know to improve choices.

### 4.3 Nudging without Unanimity

Let us now fix an arbitrary environment $E$ and consider the case where $q < 1$. To provide another analogy, reducing the requirement $q$ is tantamount to selecting between Pareto efficient policies by means of a qualified majority of $q$. As the following example illustrates, this can help even for models with the frame-cancellation property, where all frames are always Pareto efficient.

**Example 3 continued.** Reconsider the behavioral model depicted in Figure 1, assuming this time that $E$ is the set of all possible choice sets. Suppose we observe $\Lambda = \{(\succeq_3, f_1)\}$ so that $P(\Lambda) = \{\succeq_3, \succeq_4, \succeq_5\}$. The frame-cancellation property holds restricted to this set of three indistinguishable preferences, and hence we can think of it as being nested in a different model that has the frame-cancellation property globally. It follows immediately
that none of the three frames is dominated for $q = 1$, because each one is the unique non-distorting frame for one of the three possible welfare preferences.

Now let us compare frames $f_1$ and $f_2$, which are non-distorting for welfare preferences $\succeq_3$ and $\succeq_4$, respectively. For preference $\succeq_5$, frame $f_1$ induces decision mistakes between $a$ and $d$ and between $b$ and $d$. Frame $f_2$ induces these mistakes as well, plus an additional mistake between $b$ and $a$. Hence $\succeq_5$ belongs to the category of preferences for which $f_1$ induces better choices than $f_2$, which implies $r_\Lambda(f_1, f_2) = 2/3$. Analogous arguments show that $r_\Lambda(f_1, f_3) = 2/3$. For any requirement $q \leq 2/3$, we thus obtain that $f_1$ dominates both $f_2$ and $f_3$ and is an optimal nudge. While $f_2$ and $f_3$ still cannot be compared, due to $r_\Lambda(f_2, f_3) = r_\Lambda(f_3, f_2) = 1/3$, we have a clear guidance to use frame $f_1$ once we are willing to reduce our majority threshold sufficiently.

The set of indistinguishable preferences $\{\succeq_3, \succeq_4, \succeq_5\}$ in the example is asymmetric in a particular way. Preferences $\succeq_4$ and $\succeq_5$ are mutual opposites within the set, because they differ on more pairs of alternatives from each other than from the third element $\succeq_3$. An asymmetry arises because there exists no such opposite to $\succeq_3$ in the set. This is the reason why it violates the following property.

**Definition 5** A subset of preferences $\bar{P} \subseteq P$ is balanced if there exists a bijective mapping $o : \bar{P} \rightarrow P$ such that $(\succeq \setminus \succeq') \subseteq (\succeq \setminus o(\succeq))$ for all $\succeq, \succeq' \in \bar{P}$.

In a balanced set, each preference $\succeq$ can be mapped into a distinct opposite $o(\succeq)$, so that when any $\succeq' \in \bar{P}$ differs from $\succeq$ with respect to the ranking of two alternatives, then $o(\succeq)$ also differs from $\succeq$ with respect to the ranking of these two alternatives. Balance is a symmetry property which captures that the set is not tilted towards any of its elements.

Our next result will rely on the requirement that, for each welfare preference $\succeq \in P$, the set of indistinguishable preferences

$$P_\succeq = \{\succeq' \mid d(\succeq', f) = d(\succeq, f), \forall f \in F\}$$

is balanced. Many interesting behavioral models have this property. Most importantly, it holds for the entire class of limited sensitivity models introduced earlier. The set $P_\succeq$ is given by $P^c_\succeq$ in a limited sensitivity model, the set of all preferences which induce the same ordered partition of $X$ as $\succeq$. For any $\succeq' \in P^c_\succeq$, we can define its opposite $o(\succeq')$ as the preference that has the exact reverse ranking of $\succeq'$ within each block of the partition, but agrees with $\succeq'$ in the ranking of the blocks. It is easy to see that $o(\succeq') \in P^c_\succeq$ and that this mapping $o$ is bijective. The requirement of Definition 5 is satisfied because differences

---

20 Even though we often represent preferences as rankings like $b \succ a \succ d \succ c$, we remind ourselves that technically a preference $\succeq$ is a subset of the set of ordered pairs $X \times X$, so we can perform set operations. Hence $\succeq \setminus \succeq'$ is simply the set of pairs which are ranked differently by $\succeq'$ and $\succeq$. 

---
between any two preferences in $P_{\succeq}$ can arise only for alternatives within the same block, and all these differences indeed arise between $\succeq'$ and $o(\succeq')$.  

**Proposition 5** If $d$ has the frame-cancellation property and each set $P_{\succeq}$ is balanced, then the binary nudging relation $N_{q,E}(\Lambda)$ is empty for all $q, E$ and $\Lambda$.

Under the additional balance property, our previous impossibility result remains valid even if we weaken the majority requirement $q$ arbitrarily. The natural symmetry of models like satisficing or default bias – both of them special cases of limited sensitivity models and therefore covered by the result – precludes any possibility of learning enough to nudge a majority of welfare preferences into better choices.

## 5 Related Literature

The literature on behavioral welfare economics distinguishes between model-based and model-free approaches (see Bernheim, 2009; Manzini and Mariotti, 2014). The most prominent example of the model-free approach is due to Bernheim and Rangel (2009). Their purely choice-based concept has the big benefit of not requiring any conjecture about the decision process. However, as Bernheim and Rangel (2009, p. 62) have already pointed out, nudging is impossible based solely on their unambiguous choice relation without any additional assumptions or theories about decision mistakes.

In this paper, we have therefore decided to follow the model-based approach. Manzini and Mariotti (2014) distinguish between three types of problems that such an approach can encounter. Type 1 is uncertainty about the correct behavioral model. Type 2 is due to the possibility of reinterpreting the model ingredients. In our context, this for instance shows up in the question whether some dimension of the environment is welfare-relevant or part of the welfare-neutral frame. Finally, a problem of type 3 arises when multiple welfare preferences are consistent with the model and the data. Using this terminology, we have ruled out all type 1 and type 2 problems, by assuming that a behavioral model and its

---

21 There are models beyond limited sensitivity for which each set of indistinguishable preferences is balanced. Just note that any set of two preferences is balanced, so that the property holds whenever each $P_{\succeq}$ contains two elements. The property is also independent of the frame-cancellation property, as it does not preclude that frames map $P_{\succeq}$ outside of itself. Combined with the frame-cancellation property, balance is a sufficient condition for the following Proposition 5 to hold. Netzer (2017) provides other conditions that would likewise be sufficient, which are applicable when the sets of indistinguishable preferences form permutation subgroups.

22 Another interesting approach is due to Apesteguia and Ballester (2015), who propose using as a welfare benchmark the preference that is closest to a given behavior, as measured by their “swaps” criterion. Their framework does not allow for frames, but it would be interesting to develop the respective generalization and derive the implications for nudging.

23 Bernheim et al. (2015) derive weak generalized Pareto optimal $401(k)$ defaults in the sense of Bernheim and Rangel (2009), allowing for “pruning” of choices that are considered mistakes and thus deemed welfare-irrelevant.
interpretation are given, and have focussed exclusively on informational problems of type 3. Our possibility results – such as the generic possibility to identify an optimal nudge when the number of alternatives is large – must therefore be treated with some caution. Our impossibility results – such as the impossibility to identify even single dominated frames for models with the frame-cancellation property – become all the stronger. Here, the absence of type 1 and type 2 problems implies a systematic type 3 problem for nudging. The reason is that framing effects disguise precisely those aspects of welfare that would have to be known for an improvement of the decision quality by framing. Manzini and Mariotti (2014) argue that the use of stochastic choice data can sometimes mitigate the type 3 problem. We leave an exploration of this possibility to future research.

Goldin and Reck (2015) also study the problem of identifying welfare preferences when choices are distorted by frames, focussing mostly on binary choice problems with defaults. They estimate the preference shares among fully rational agents by the shares of agents who choose each alternative when it is not the default. The preference shares among the inconsistent agents are then deduced under identifying assumptions, for instance the assumption that they are identical to the rational agents after controlling for observable differences. It is then possible to identify the default that induces the best choice for a majority of the population.

Informational requirements are not the only obstacle that a libertarian paternalist has to overcome. Spiegler (2015) emphasizes that equilibrium reactions by firms must be taken into account when assessing the consequences of a nudge-based policy. Even abstracting from informational problems, these reactions can wipe out the intended benefits of a policy. By contrast, Jimenez-Gomez (2016) argues that nudging is still welfare-increasing when competition is perfect.

Finally, frames are often not chosen by a benevolent regulator but by profit-maximizing actors in markets, which also gives rise to questions about welfare. Siegel and Salant (2015) study contracts when a seller is able to temporarily influence the buyers’ willingness to pay by framing. They provide conditions under which optimal contracts make use of strategic framing, show how framing interacts with market regulation, and discuss the welfare implications.

6 Conclusions

We have taken the revealed preference perspective for a single agent. Aside from its methodological justification, this is also directly relevant for nudging, where “personalization does appear to be the wave of the future” (Sunstein, 2014, p. 100). In the digital age of big data, individual-specific data gathering and nudging is achievable, for instance by relying on cookies. However, our results also speak to the problem of nudging a popula-
tion of agents. On the elicitation stage, an assumption that different agents have identical preferences, possibly after controlling for observables, or are drawn representatively from a population, would allow us to combine observations of different agents into a single data set, facilitating the preference elicitation. On the nudging stage, the necessity to determine one frame for a population of heterogeneous agents gives rise to ordinary social choice problems, which we have mostly refrained from studying in this paper.

At first glance, similar problems to the ones we have studied here should arise if the frame is not chosen by a benevolent regulator but by a profit-maximizing firm. A second glance reveals substantial differences. When framing its offer, the profit-maximizing firm will want to take the consumer's behavioral reaction into account, but it does not care about the consumer's welfare preference per se. As a consequence, some of the issues that we have documented will not arise. As an example, consider a monopolist whose product portfolio is given by $X$. If the consumer is a perfect-recall satisficer who wants to buy exactly one product, the best the monopolist can do is to present the products in order of their profitability. Whenever the frame has an effect at all (which depends on the consumer's welfare preference, aspiration level, and budget set), it manipulates her into buying the product that yields the largest profits. No information about welfare preferences is necessary to determine this optimal marketing strategy. Exploiting the behavioral agent is a quite different problem from helping her make better decisions.

Our analysis shows that seemingly small differences between decision processes can make a big difference for nudging. For instance, a satisficing agent stops searching as soon as some aspiration level is achieved. Our results imply that it is impossible to help this agent make systematically better choices. If the agent stops searching at the end of a search engine's result page, by contrast, it is relatively easy to improve her choices by framing. A major difference between these two processes seems to be that stopping is endogenous to the welfare preference in the first case and exogenous in the second. Observations like this raise important questions for future research on decision processes.

Finally, there are natural extensions of our framework worth studying in future work. Our approach has been non-Bayesian, in the same way the ordinary revealed preference approach is not Bayesian. It would be interesting to see how probabilistic prior knowledge of the desired behavior interacts with learning by the revealed preference logic. Moreover, the assumption that framing is the only reason for incoherence of choice deserves to be dropped. There are interesting decision procedures which generate violations of rationality axioms already when a frame is fixed (e.g. Masatlioglu and Nakajima, 2015). The regulator is then confronted with the additional problem of having to compare the different behavioral preference reversals that different frames induce.
References


A Proofs

A.1 Proof of Proposition 1

We first provide a useful general characterization of the binary relation \( N(\Lambda) \). It relies on the sets of ordered pairs \( B(\succeq, f) = d(\succeq, f) \setminus \succeq \), which record all binary comparisons that are reversed from \( \succeq \) by \( f \).

**Lemma 1** For any \( \Lambda \) and \( [f]_\Lambda \neq [f']_\Lambda \), it holds that \( [f]_\Lambda N(\Lambda)[f']_\Lambda \) if and only if

\[ B(\succeq, f) \subseteq B(\succeq, f') \text{ for each } \succeq \in P(\Lambda). \]

**Proof.** Suppose first that \( B(\succeq, f) \subseteq B(\succeq, f') \) holds for all \( \succeq \in P(\Lambda) \). To show that this implies \( [f]_\Lambda N(\Lambda)[f']_\Lambda \), we proceed by contradiction and assume that there exist \( \succeq \in P(\Lambda) \) and \( S \subseteq X \) for which \( c(d(\succeq, f), S) = x \) and \( c(d(\succeq, f'), S) = y \) with \( x \neq y \) and \( y \succeq x \). The definition of \( c \) implies \((x, y) \in d(\succeq, f)\) and \((x, y) \notin d(\succeq, f')\). Together with \((x, y) \notin \succeq \) this implies \((x, y) \in B(\succeq, f)\) but \((x, y) \notin B(\succeq, f')\), a contradiction.

For the converse, suppose that there exist \( \succeq \in P(\Lambda) \) and \( x, y \in X \) with \((x, y) \in B(\succeq, f)\) but \((x, y) \notin B(\succeq, f')\), which requires \( x \neq y \). This implies \((x, y) \in d(\succeq, f)\) and \((x, y) \notin d(\succeq, f')\). Then \( c(d(\succeq, f'), \{x, y\}) = y \geq x = c(d(\succeq, f), \{x, y\}) \), which implies that \( [f]_\Lambda N(\Lambda)[f']_\Lambda \) does not hold. \( \blacksquare \)

For the if-statement, suppose \( \succeq \) is identifiable, which implies that \( \bar{\Lambda}(\succeq) \) is not identical to \( \bar{\Lambda}(\succeq') \) for any other \( \succeq' \). Then \( P(\bar{\Lambda}(\succeq)) = \{\succeq\} \). Consider any \( f \) with \( d(\succeq, f) = \succeq \), which exists by assumption. For any \( f' \in F \), we then have \( B(\succeq, f) = \emptyset \subseteq B(\succeq, f') \). Hence, by Lemma 1, we obtain \( [f]_{\Lambda(\succeq)} N(\bar{\Lambda}(\succeq))[f']_{\Lambda(\succeq)} \) for all \( [f']_\Lambda \neq [f]_\Lambda \), which implies \( f \in G(\bar{\Lambda}(\succeq)) \).

For the only-if statement, suppose that \( \succeq \) is not identifiable, i.e., there exists \( \succeq' \neq \succeq \) with \( \bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq) \). Then \( \{\succeq, \succeq'\} \subseteq P(\bar{\Lambda}(\succeq)) \). Consider any \( f_1 \) with \( d(\succeq, f_1) = \succeq \) and any \( f_2 \) with \( d(\succeq', f_2) = \succeq' \), so that \( B(\succeq, f_1) = \emptyset \) and \( B(\succeq', f_2) = \emptyset \). Assume by contradiction that there exists some \( f \in G(\bar{\Lambda}(\succeq)) \). Then \( B(\succeq, f) = \emptyset \) must hold, as otherwise \( [f]_{\Lambda(\succeq)} \neq [f_1]_{\bar{\Lambda}(\succeq)} \) but not \( [f]_{\Lambda(\succeq)} N(\bar{\Lambda}(\succeq))[f_1]_{\bar{\Lambda}(\succeq)} \) according to Lemma 1, a contradiction to \( f \) being an optimal nudge. The same argument for \( f_2 \) also implies that \( B(\succeq', f) = \emptyset \) must also hold. Hence \( d(\succeq, f) = \succeq \) and \( d(\succeq', f) = \succeq' \), which contradicts that \( \bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq) \).

A.2 Proof of Proposition 2

Any behavioral model \( d \) is characterized by the collection of maximal data sets \( (\bar{\Lambda}(\succeq))_{\succeq \in P} \) that it assigns to the welfare preferences. Suppose there are \( m_P \geq 2 \) preferences and
$m_F \geq 2$ frames. Then there are $(m_P)^{m_F}$ different maximal data sets. For a given welfare preference $\succeq$, however, only

$$N(m_P, m_F) = (m_P)^{m_F} - (m_P - 1)^{m_F}$$

of them are admissible, as the others contradict the existence of a non-distorting frame for $\succeq$. The number of possible models is thus given by $N(m_P, m_F)^{m_P}$. To obtain a model with identifiable preferences, we need to assign a different maximal data set to each welfare preference. Suppose we assign one of the $N(m_P, m_F)$ admissible data sets to the first welfare preference. Then there remain at least $N(m_P, m_F) - 1$ admissible data sets for the second welfare preference (the exact number is still $N(m_P, m_F)$ if the data set assigned to the first preference was not admissible for the second preference), and so on. Observe that $N(m_P, m_F) \geq m_P$, so we can proceed iteratively and obtain the falling factorial

$$N(m_P, m_F)^{m_P} = N(m_P, m_F) \times (N(m_P, m_F) - 1) \times \ldots \times (N(m_P, m_F) - m_P + 1)$$

as a lower bound on the number of models with identifiable preferences. Consequently,

$$S(m_P, m_F) = \frac{N(m_P, m_F)^{m_P}}{N(m_P, m_F)^{m_P}}$$

is a lower bound on the share of models with identifiable preferences. We can rewrite

$$S(m_P, m_F) = \frac{N(m_P, m_F)^{m_P}}{N(m_P, m_F)^{m_P}} \times \frac{N(m_P, m_F) - 1}{N(m_P, m_F)} \times \ldots \times \frac{N(m_P, m_F) - m_P + 1}{N(m_P, m_F)}$$

$$= \prod_{k=1}^{m_P-1} \left(1 - \frac{k}{N(m_P, m_F)}\right)$$

$$= \exp\left(\sum_{k=1}^{m_P-1} \log\left(1 - \frac{k}{N(m_P, m_F)}\right)\right),$$

where $1 > k/N(m_P, m_F) > 0$ holds for all $k = 1, \ldots, m_P - 1$. Recall that for $x > -1$ we have $\log(1 + x) \geq x/(1 + x)$, which implies

$$\sum_{k=1}^{m_P-1} \log\left(1 - \frac{k}{N(m_P, m_F)}\right) \geq \sum_{k=1}^{m_P-1} -\frac{k}{N(m_P, m_F) - k}.$$
Furthermore,

\[
\sum_{k=1}^{m_P-1} \frac{k}{N(m_P, m_F) - k} \geq \sum_{k=1}^{m_P-1} \frac{k}{N(m_P, m_F) - m_P + 1} = \frac{1}{N(m_P, m_F) - m_P + 1} \sum_{k=1}^{m_P-1} k = \frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)}.
\]

Altogether, we therefore have

\[S(m_P, m_F) \geq \exp \left( -\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)} \right) = \tilde{S}(m_P, m_F),\]

so \(\tilde{S}(m_P, m_F)\) is also a lower bound on the share of models with identifiable preferences.

We are interested in asymptotic behavior as the number of alternatives \(m_X\) and hence the number of preferences \(m_P(m_X) = m_X!\) grows. Holding \(m_F\) fixed and treating \(m_P\) as a real variable, it follows with l’Hôpital’s rule that

\[
\lim_{m_P \to \infty} -\frac{(m_P)^2 - m_P}{2(N(m_P, m_F) - m_P + 1)} = 0
\]

whenever \(m_F \geq 4\). We thus obtain \(\lim_{m_X \to \infty} \tilde{S}(m_P(m_X), m_F) = 1\) whenever \(m_F \geq 4\).

Now consider the case that the number of frames \(m_F(m_X)\) also depends on the number of alternatives. Observe that \(\tilde{S}(m_P, m_F)\) is strictly increasing in \(m_F\) whenever \(m_P \geq 2\). At the same time, \(\tilde{S}(m_P, m_F) \leq 1\) always holds since \(\tilde{S}(m_P, m_F)\) is a lower bound on a proportion. Hence we obtain

\[
\lim_{m_X \to \infty} \tilde{S}(m_P(m_X), m_F(m_X)) = 1
\]

whenever there exists \(m\) such that \(m_F(m_X) \geq 4\) for all \(m_X \geq m\). Then the share of models with identifiable preferences converges to 1 as the number of alternatives grows to infinity.

**A.3 Proof of Proposition 3**

Consider any \(d\) with the frame-cancellation property and any data set \(\Lambda\). Fix any two frames \(f_1, f_2 \in F\) such that \([f_1]_\Lambda \neq [f_2]_\Lambda\). Then there exists \(\succeq \in P(\Lambda)\) such that

\[d(\succeq, f_1) = \succeq_1 \neq \succeq_2 = d(\succeq, f_2).\]
By the frame-cancellation property, we have \(d(\succeq_1, f) = d(d(\succeq, f_1), f) = d(\succeq, f)\) for all \(f \in F\), which implies that \(\succeq_1 \in P(\Lambda)\). Furthermore, it implies \(d(\succeq_1, f_1) = d(\succeq, f_1) = \succeq_1\) and \(d(\succeq_1, f_2) = d(\succeq, f_2) = \succeq_2\), so that \(B(\succeq_1, f_1) = \emptyset\) and \(B(\succeq_1, f_2) \neq \emptyset\). Hence

\[
B(\succeq_1, f_1) \subset B(\succeq_1, f_2).
\]

The same argument applied to \(\succeq_2\) yields \(\succeq_2 \in P(\Lambda)\) and

\[
B(\succeq_2, f_2) \subset B(\succeq_2, f_1).
\]

Lemma 1 now implies that neither \([f_1]_\Lambda N(\Lambda)[f_2]_\Lambda\) nor \([f_2]_\Lambda N(\Lambda)[f_1]_\Lambda\).

### A.4 Proof of Proposition 4

Consider any \(d\) with the frame-cancellation property, any environment \(E\), and any data set \(\Lambda\). Fix any two frames \(f_1, f_2 \in F\) such that \([f_1]_\Lambda \neq [f_2]_\Lambda\). Then there exist \(\succeq \in P(\Lambda)\) and \(S' \in E\) such that

\[
c(d(\succeq, f_1), S') = x_1 \neq x_2 = c(d(\succeq, f_2), S').
\]

Denote \(\succeq_1 = d(\succeq, f_1)\) and \(\succeq_2 = d(\succeq, f_2)\), so that \(x_1 \succeq_1 x_2\) and \(x_2 \succeq_2 x_1\) holds. The frame-cancellation property implies \(d(\succeq_1, f) = d(d(\succeq, f_1), f) = d(\succeq, f)\) for all \(f \in F\), so that \(\succeq_1 \in P(\Lambda)\). Analogously, \(d(\succeq_2, f) = d(d(\succeq, f_2), f) = d(\succeq, f)\) for all \(f \in F\), which implies \(\succeq_2 \in P(\Lambda)\). We then obtain

\[
c(d(\succeq_1, f_1), S') = c(d(\succeq_1, f_1), S') = x_1 \succeq_1 x_2 = c(d(\succeq, f_2), S') = c(d(\succeq_1, f_2), S')
\]

and

\[
c(d(\succeq_2, f_2), S') = c(d(\succeq_2, f_2), S') = x_2 \succeq_2 x_1 = c(d(\succeq_1, f_1), S') = c(d(\succeq_2, f_1), S').
\]

This implies \(r_\Lambda(f_1, f_2) < 1\) and \(r_\Lambda(f_2, f_1) < 1\) (where \(r_\Lambda(f_2, f_1)\) denotes the share of preferences for which \(f_2\) outperforms \(f_1\)). Hence we have neither \([f_1]_\Lambda N_{1,E}(\Lambda)[f_2]_\Lambda\) nor \([f_2]_\Lambda N_{1,E}(\Lambda)[f_1]_\Lambda\).

### A.5 Proof of Proposition 5

Consider any \(d\) with the frame-cancellation property and balanced sets \(P_{\succeq}\), and any environment \(E\). Suppose first that \(\Lambda\) is non-empty, so there exists \((\succeq', f') \in \Lambda\). We claim that \(P(\Lambda) = P_{\succeq'}\). To see why, suppose \(\succeq \in P(\Lambda)\), so that \(d(\succeq, f') = \succeq'\). Then the frame-cancellation property implies \(d(\succeq', f) = d(d(\succeq, f'), f) = d(\succeq, f)\) for all \(f \in F\). Hence
we have $\succeq \in P_{\succeq'}$, which implies $P(\Lambda) \subseteq P_{\succeq'}$. We also obtain $\succeq' \in P(\Lambda)$, which implies $P_{\succeq'} \subseteq P(\Lambda)$. In particular, we can now conclude that $c(d(\succeq, f), S) = c(d(\succeq', f), S)$ holds for all $\succeq \in P(\Lambda) = P_{\succeq'}$ and all $f \in F$ and $S \in E$.

Fix any two frames $f_1, f_2 \in F$ such that $[f_1]_\Lambda \neq [f_2]_\Lambda$. Denoting $\succeq_1 = d(\succeq', f_1)$ and $\succeq_2 = d(\succeq', f_2)$, we again obtain from the frame-cancellation property that $\succeq_1, \succeq_2 \in P(\Lambda) = P_{\succeq'}$. Furthermore, there exists $S' \in E$ such that

$$c(d(\succeq', f_1), S') = c(\succeq_1, S') = c(\succeq_2, S') = c(d(\succeq', f_2), S').$$

Suppose $f_1$ strictly outperforms $f_2$ for some $\succeq \in P(\Lambda) = P_{\succeq'}$, in the sense that

$$c(d(\succeq, f_1), S) = c(\succeq_1, S) \succeq c(\succeq_2, S) = c(d(\succeq, f_2), S)$$

for all $S \in E$. We claim that $f_2$ then strictly outperforms $f_1$ for $o(\succeq) \in P(\Lambda) = P_{\succeq'}$, where $o(\succeq)$ is the opposite of $\succeq$ in the balanced set $P_{\succeq'}$. Consider any $S \in E$ for which $c(\succeq_1, S) = x_1 \neq x_2 = c(\succeq_2, S)$. Then $(x_1, x_2) \in \succeq$, $(x_1, x_2) \in \succeq_1$ and $(x_2, x_1) \in \succeq_2$. Therefore $(x_1, x_2) \in (\succeq \setminus \succeq_2)$, so $(x_1, x_2) \in (\succeq \setminus o(\succeq))$ by definition of the mapping $o$. It follows that $(x_2, x_1) \in o(\succeq)$, or

$$c(d(o(\succeq), f_2), S) = c(\succeq_2, S) = x_2 o(\succeq) x_1 = c(\succeq_1, S) = c(d(o(\succeq), f_1), S),$$

which implies our claim.

Bijectiveness of the mapping $o$ now implies $r_\Lambda(f_1, f_2) = r_\Lambda(f_2, f_1) \leq 1/2$, so we have neither $[f_1]_\Lambda N_{q,E}(\Lambda)[f_2]_\Lambda$ nor $[f_2]_\Lambda N_{q,E}(\Lambda)[f_1]_\Lambda$, for any $q > 1/2$.

The proof for $\Lambda = \emptyset$ follows immediately from the fact that the number of welfare preferences for which $f_1$ outperforms $f_2$ is the same as the number of welfare preferences for which $f_2$ outperforms $f_1$ already in each set $P_{\succeq}$ separately, and $\{P_{\succeq} \mid \succeq \in P\}$ forms a partition of $P(\emptyset) = P$. 
B Additional Material

B.1 Model Uncertainty

In the main text, we assumed that there is a unique conjecture about the behavioral model, while it may be more appropriate to assume that the regulator considers a number of different models possible. We can replace the assumption of a unique behavioral model by the assumption that the regulator considers any distortion function \( d \in D \) possible, where \( D \) is a given set of conjectures. For instance, there could be uncertainty about the aspiration level of a satisficer, or one of the models in \( D \) could be the rational agent.\(^{24}\) As a consequence, we no longer have to learn about the welfare preference only, but about the pair \((d, \succeq) \in D \times P\) of the distortion function and the welfare preference.\(^{25}\)

Let \( \bar{\Lambda}(d, \succeq) = \{(d(\succeq, f), f) \mid f \in F\} \) denote the maximal data set generated by the pair \((d, \succeq)\). Then the set of pairs \((d, \succeq)\) that are consistent with data set \( \Lambda \) is \( DP(\Lambda) = \{(d, \succeq) \mid \Lambda \subseteq \bar{\Lambda}(d, \succeq)\} \). We again assume that \( DP(\Lambda) \) is non-empty, i.e., at least one conjecture is not falsified by the data. Once we have narrowed down the set of model-preference pairs to \( DP(\Lambda) \), we obtain the equivalence class of frame \( f \) by \( [f]_\Lambda = \{f' \mid d(\succeq, f) = d(\succeq, f'), \forall (d, \succeq) \in DP(\Lambda)\} \). We can then modify our definition of the binary nudging relation in a natural way, taking into account that both model and welfare preference are unknown. In particular, for any \([f]_\Lambda \neq [f']_\Lambda\) we define \([f]_\Lambda N(\Lambda) [f']_\Lambda\) if for each \((d, \succeq) \in DP(\Lambda)\) it holds that \( c(d(\succeq, f), S) \succeq c(d(\succeq, f'), S)\) for all non-empty \( S \subseteq X\), so that for each remaining behavioral model the agent’s choices under frame \( f \) are better than under \( f'\), no matter which of the welfare preferences that are consistent with the behavioral model and the data set is the true one.

We are again interested in the existence of an optimal nudge. By the same reasoning as in Section 3, we consider maximal data sets only. An immediate extension of Definition 2 could require identifiability of \( \succeq \) in \( d \), for a given pair \((d, \succeq)\). This property is in fact necessary but no longer sufficient for the existence of an optimal nudge. It rules out that the maximal data set \( \bar{\Lambda}(d, \succeq) \) could have been generated by a different welfare preference \( \succeq' \) and the same model \( d \), but it does not rule out that it could have been generated by a different welfare preference \( \succeq' \) and a different model \( d' \). Since two behaviorally equivalent model-preference pairs \((d, \succeq)\) and \((d', \succeq')\) can have different normative implications (see e.g. Kőszegi and Rabin, 2008b; Bernheim, 2009; Masatlioglu et al., 2012), identifiability in the extended setting must aim at all aspects of the pair \((d, \succeq)\) that are normatively relevant.

\(^{24}\)It is central to the idea of asymmetric paternalism (Camerer et al., 2003) that there are different types of agents, some of which are rational and should not be restricted by regulation.

\(^{25}\)We continue to assume that there is a non-distorting frame for each pair \((d, \succeq)\), which will typically depend both on the model and on the welfare preference.
Definition 6 Pair \( (d, \succeq) \) is virtually identifiable if for each \( (d', \succeq') \in D \times P \) with \( \succeq' \neq \succeq \), there exists \( f \in F \) such that \( d(\succeq, f) \neq d'(\succeq', f) \).

Virtual identifiability implies that the welfare preference \( \succeq \) is known for sure once the maximal data set has been collected. It still allows for some uncertainty about the behavioral model, but only to the extent that we may not be able to predict the behavior of an agent with a different welfare preference \( \succeq' \neq \succeq \).

Proposition 6 With model uncertainty, \( G(\bar{\Lambda}(d, \succeq)) \) is non-empty if and only if \( (d, \succeq) \) is virtually identifiable.

Proof. The proof is similar to the proof of Proposition 1 and therefore omitted. ■

We can have multiple models with identifiable preferences each, that, if considered jointly, do not have virtually identifiable model-preference pairs. Model uncertainty of this type poses a fundamental new problem to nudging. On the other hand, adding a rational agent to any given behavioral model with identifiable preferences always preserves the property of virtually identifiable model-preference pairs.

B.2 Imperfectly Observable Frames

In the main text, we assumed that frames are perfectly observable and controllable by the regulator. Since a frame can be very complex, this assumption deserves to be relaxed. The generalization also allows us to model internal states that affect the agent’s choices. For instance, consider a satisficing model in which the aspiration level \( k \) fluctuates in a non-systematic and unobservable way, as in the original RS model. We can capture this by including the aspiration level into the frame (\( k \) affects choice but not welfare), but the extended frame cannot be fully observable and controllable for an outsider.

Imperfect observability can be modelled as a structure \( \Phi \subseteq 2^F \) with the property that for each \( f \in F \) there exists \( \phi \in \Phi \) with \( f \in \phi \). The interpretation is that the regulator observes only sets of frames \( \phi \in \Phi \) and does not know under which of the frames \( f \in \phi \) the agent was acting. The example with a fluctuating aspiration level can be modelled as \( F = P \times \{2, \ldots, m_X\} \) and \( \Phi = \{\phi_p \mid p \in P\} \) for \( \phi_p = \{(p, k) \mid k \in \{2, \ldots, m_X\}\} \).

A behavioral data set is a subset \( \Lambda \subseteq P \times \Phi \), where \( (\succeq', \phi') \in \Lambda \) means that the agent has been observed behaving according to \( \succeq' \) when the frame must have been one of the elements of \( \phi' \). Thus a welfare preference \( \succeq \) is consistent with \( \Lambda \) if for each \( (\succeq', \phi') \in \Lambda \) we have \( \succeq' = d(\succeq, f') \) for some \( f' \in \phi' \), so that \( \succeq \) might have generated the data set from the regulator’s perspective. The set of welfare preferences that are consistent with \( \Lambda \) is \( P(\Lambda) = \{\succeq \mid \Lambda \subseteq \bar{\Lambda}(\succeq)\} \), where \( \bar{\Lambda}(\succeq) = \{(d(\succeq, f), \phi) \mid f \in \phi \in \Phi\} \) is again the maximal data set for \( \succeq \). Note that a non-singleton set of frames \( \phi \) can appear more than once.
in a maximal data set, combined with different behavioral preferences. This also implies
that the cardinality of $\bar{\Lambda}(\succeq)$ is no longer the same for all $\succeq \in P$, because two different
frames $f, f' \in \phi$ might generate two different observations for some preference but only
one observation for another preference.

In many applications, such as a satisficing model with fluctuating aspiration level, it
is reasonable to assume that the same $\Phi$ applies to observing and nudging, i.e., the frame
dimensions that the regulator can observe are identical to those that he can control. We
allow for the more general case where a set of frames can be chosen as a nudge from a
potentially different structure $\Phi_N$. When comparing two elements $\phi, \phi' \in \Phi_N$, we will
not necessarily want to compare the agents’ choices under each $f \in \phi$ with her choices
under each $f' \in \phi'$. For instance, we want to compare orders of presentation for each
aspiration level separately, not across aspiration levels. To this end, we introduce a set $H$
of selection functions, which are functions $h : \Phi_N \to F$ with the property that $h(\phi) \in \phi$.
The elements of $H$ capture the comparisons that we need to make: when comparing $\phi$
with $\phi'$ we compare only the choices under the frames $h(\phi)$ and $h(\phi')$, for each $h \in H$. In
the satisficing model we would have one $h_k \in H$ for each aspiration level $k \in \{2, \ldots, m_X\}$,
defined by $h_k(\phi_p) = (p, k)$. The only assumption that we impose on $H$ is that for each
$f \in \phi \in \Phi_N$ there exist $h \in H$ such that $h(\phi) = f$. We can then define the equivalence
class $[\phi]_\Lambda = \{\phi' \mid d(\succeq, h(\phi')) = d(\succeq, h(\phi)), \forall (h, \succeq) \in H \times P(\Lambda)\}$ for any $\Lambda$ and $\phi$. As
before, for any $[\phi]_\Lambda \neq [\phi']_\Lambda$, let $[\phi]_\Lambda N(\Lambda)[\phi']_\Lambda$ if for each $(h, \succeq) \in H \times P(\Lambda)$ it holds that
c($d(\succeq, h(\phi)), S) \geq c(d(\succeq, h(\phi')), S)$, for all non-empty $S \subset X$.

Let $G(\Lambda) = \{\phi \mid [\phi]_\Lambda N(\Lambda)[\phi']_\Lambda \text{ for all } [\phi']_\Lambda \neq [\phi]_\Lambda\}$ be the set of optimal nudges. We
again consider maximal data sets. An immediate extension of Definition 2 could require
that for each $\succeq' \neq \succeq$ there exists $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f)$. This property
turns out to be necessary but not sufficient for $G(\bar{\Lambda}(\succeq))$ to be non-empty. It implies
that the maximal data set for $\succeq$ is different from the maximal data set for every other
preference, so that $\succeq$ is identified once $\bar{\Lambda}(\succeq)$ has been collected and once it is known that
this set is indeed maximal. Unfortunately, the cardinality of $\bar{\Lambda}(\succeq)$ no longer carries that
kind of information, as we could have $\bar{\Lambda}(\succeq) \subset \bar{\Lambda}(\succeq')$ for some $\succeq' \neq \succeq$. Upon observing
$\bar{\Lambda}(\succeq)$ we then never know if we have already arrived at the maximal data set for $\succeq$, or
if there is an additional observation yet to be made. Our notion of identifiability in the
setting with imperfectly observable frames must therefore ensure that the maximal data
set reveals itself as maximal.

---

26In continuation of our previous approach, we assume that for each $\succeq \in P$ there exists $\phi \in \Phi_N$ such
that $d(\succeq, f) = \succeq$ for all $f \in \phi$. This implies that nudging is not per se impeded by the lack of control over
frames. The assumption is clearly much stronger here than before. For instance, it holds in the described
satisficing application when there is perfect recall (because the order of presentation that coincides with
the welfare preference is non-distorting for all possible aspiration levels) but would not hold with no recall
(because the non-distorting order of presentation then depends on the aspiration level).
Definition 7  Welfare preference $\succeq$ is potentially identifiable if for each $\succeq' \in P$ with $\succeq' \neq \succeq$, there exist $f \in \phi \in \Phi$ such that $d(\succeq, f) \neq d(\succeq', f')$ for all $f' \in \phi$.

When frames are not directly observed, identifiability requires more than the existence of a frame $f \in \phi \in \Phi$ that distinguishes between $\succeq$ and $\succeq'$. We can exclude welfare preference $\succeq'$ as a candidate only if the observed distorted preference $d(\succeq, f)$ could not as well have been generated by $\succeq'$ for any other $f' \in \phi$. For instance, no preference is potentially identifiable in the perfect-recall satisficing model with fluctuating aspiration level.\(^{27}\)

Proposition 7  With imperfectly observable frames, $G(\tilde{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is potentially identifiable.

Proof. The proof is similar to the proof of Proposition 1 and therefore omitted. ■

We use the term potential identifiability because there is no guarantee that we will ever be able to collect $\tilde{\Lambda}(\succeq)$. Even if the agent is exposed repeatedly to a set of frames $\phi$, it can still happen that a specific element $f \in \phi$ does not materialize. This is in contrast to the case of observable frames, where a maximal data set can always be collected in exactly $m_F$ steps.

B.3 Complexity

In this appendix, we focus attention on models with identifiable welfare preferences, for which knowledge of an optimal nudge is guaranteed once a maximal data set has been observed. However, collecting a maximal data set requires observing the agent under all $m_F$ frames, which might be beyond the regulator’s means. We are thus interested in optimal data gathering procedures and the required quantity of information. The idea is that a regulator, who ultimately seeks to impose the optimal nudge, is also able to impose a specific sequence of frames on the agent, with the goal of eliciting the necessary information efficiently.

For each $t \in \{0, 1, \ldots, m_F\}$, let

$$L_t = \{\Lambda | P(\Lambda) \neq \emptyset \text{ and } |\Lambda| = t\}$$

be the collection of data sets that do not falsify the behavioral model and contain exactly $t$ observations, i.e., observations for $t$ different frames. In particular, $L_0 = \{\emptyset\}$ and $L_{m_F}$

\(^{27}\)To see why, note that two preferences which coincide except for the ranking of the two top alternatives are behaviorally equivalent for every order of presentation and every aspiration level $k \geq 2$. This was different if we allowed the agent to be sometimes rational ($k = 1$) as in the original RS model, in which case all preferences are potentially identifiable.
consists of all maximal data sets, i.e., those data sets that the regulator may end up with after observing the agent under all possible frames. Then $L = L_0 \cup L_1 \cup \ldots \cup L_{m_F-1}$ is the collection of all possible data sets except the maximal ones. An elicitation procedure dictates for each of these data sets a yet unobserved frame, under which the agent is to be observed next.

**Definition 8** An elicitation procedure is a mapping $e : L \rightarrow F$ with the property that, for each $\Lambda \in L$, there does not exist $(\succeq, f) \in \Lambda$ such that $e(\Lambda) = f$.

A procedure $e$ starts with the frame $e(\emptyset)$ and, if the welfare preference is $\succeq$, generates the first data set $\Lambda_1(e, \succeq) = \{(d(\succeq, e(\emptyset)), e(\emptyset))\}$. It then dictates the different frame $e(\Lambda_1(e, \succeq))$ and generates a larger data set $\Lambda_2(e, \succeq)$ by adding the resulting observation. This yields a sequence of expanding data sets described recursively by $\Lambda_0(e, \succeq) = \emptyset$ and

$$
\Lambda_{t+1}(e, \succeq) = \Lambda_t(e, \succeq) \cup \{(d(\succeq, e(\Lambda_t(e, \succeq))), e(\Lambda_t(e, \succeq)))\},
$$

until the maximal data set $\Lambda_{m_F}(e, \succeq) = \overline{\Lambda}(\succeq)$ is reached. Hence all elicitation procedures deliver the same outcome after $m_F$ steps, but typically differ at earlier stages. A procedure does not use any exogenous information about the welfare preference, but the frame to be dictated next can depend on the information generated endogenously by the growing data set.\(^\text{28}\)

We now define the complexity $n$ of the nudging problem as the number of steps that the quickest elicitation procedure requires until it identifies an optimal nudge for sure. Formally, let

$$
n(e, \succeq) = \min \{ t \mid G(\Lambda_t(e, \succeq)) \neq \emptyset \}
$$

denote the first step at which $e$ identifies an optimal nudge if the welfare preference is $\succeq$. Since this preference is unknown, $e$ guarantees a result only after $\max_{\succeq \in \mathcal{P}} n(e, \succeq)$ steps. With $\Pi$ denoting the set of all elicitation procedures, we have to be prepared to gather

$$
n = \min_{e \in \Pi} \max_{\succeq \in \mathcal{P}} n(e, \succeq)
$$

data points before we can nudge successfully.

To illustrate the concepts, we first consider the limited search model (assuming $m_X \geq 3$ to make all preferences identifiable). The following result shows that learning and nudging are relatively simple in this model.

\(^{28}\)Notice that an elicitation procedure dictates frames also for pre-collected data sets that itself never generates. We tolerate this redundancy because otherwise definitions and proofs would become substantially more complicated, at no gain.
Proposition 8 For any \( m_X \geq 3 \), the limited search model satisfies

\[
    n = \begin{cases} 
    3 & \text{if } k = m_X/2 \text{ and } k \text{ is odd,} \\
    2 & \text{otherwise.}
    \end{cases}
\]

Proof. We assume \( k \leq m_X/2 \) throughout the proof, as cases where \( k > m_X/2 \) can be dealt with equivalently by reversing the role of the first page \( f \) and the second page \( X \setminus f \) of the search engine.

Case 1: \( k \) even. We first construct an elicitation procedure \( e \) and then show that it is optimal. Let \( e(\emptyset) = f_1 \) be an arbitrary subset \( f_1 \subseteq X \) with \( |f_1| = k \). Now fix any welfare preference \( \succeq \). The procedure then generates a data set \( \Lambda_1 = \{ (\succeq_1, f_1) \} \in L_1 \), where \( \succeq_1 \) agrees with \( \succeq \) within the sets \( f_1 \) and \( X \setminus f_1 \). Let \( a_i \) denote the alternative ranked at position \( i \) within the set \( f_1 \) by \( \succeq_1 \), for each \( i = 1, \ldots, k \). Let \( b_i \) denote the alternative ranked at position \( i \) within the set \( X \setminus f_1 \) by \( \succeq_1 \), for each \( i = 1, \ldots, k, \ldots, m_X - k \). Then construct the frame \( e(\Lambda_1) = f_2 \) as \( f_2 = \{ a_1, \ldots, a_{k/2}, b_{k/2+1}, \ldots, b_k \} \). The procedure then generates a data set \( \Lambda_2 = \{ (\succeq_1, f_1), (\succeq_2, f_2) \} \in L_2 \), where \( \succeq_2 \) agrees with \( \succeq \) within the sets \( f_2 \) and \( X \setminus f_2 \). This construction is applied to all the data sets \( \Lambda_i \) that are generated by the elicitation procedure for some welfare preference. The elicitation procedure can be continued arbitrarily for all other data sets.

Let \( \succeq \) be an arbitrary true welfare preference. We claim that the set \( T_k(\succeq) \) of top \( k \) alternatives according to \( \succeq \) can be deduced from the generated \( \Lambda_2 \), so that the optimal nudge is identified and \( n(e, \succeq) \leq 2 \) follows. Observe first that none of the alternatives \( b_{k+1}, \ldots, b_{m_X - k} \) (if they exist) can belong to \( T_k(\succeq) \), because \( \Lambda_1 \) has already revealed that each \( b_1, \ldots, b_k \) is preferred by \( \succeq \). Now suppose that \( b_k \succeq b_1 \) holds. We then know that \( b_k \succeq a_1 \) and thus \( T_k(\succeq) = \{ b_1, \ldots, b_k \} \). Otherwise, if \( a_1 \succeq b_k \) holds, we know that \( a_1 \succeq b_k \) and thus \( b_k \notin T_k(\succeq) \) but \( a_1 \in T_k(\succeq) \). In this case we can repeat the argument for \( a_2 \) and \( b_{k-1} \): if \( b_{k-1} \succeq a_2 \) we know that \( b_{k-1} \succeq a_2 \) and thus \( T_k(\succeq) = \{ b_1, \ldots, b_{k-1}, a_1 \} \); otherwise, if \( a_2 \succeq b_{k-1} \) holds, we know that \( a_2 \succeq b_{k-1} \) and thus \( b_{k-1} \notin T_k(\succeq) \) but \( a_2 \in T_k(\succeq) \). Iteration either reveals \( T_k(\succeq) \) or arrives at \( a_{k/2} \succeq b_{k/2+1} \), which implies \( a_{k/2} \succeq b_{k/2+1} \). In this case, we know that \( T_k(\succeq) \) consists of \( a_1, \ldots, a_{k/2} \) and those \( k/2 \) alternatives that \( \succeq_2 \) and hence \( \succeq \) ranks top within \( X \setminus f_2 \).

Since \( \succeq \) was arbitrary, we know that \( \max_{\succeq \in P} n(e, \succeq) \leq 2 \). Obviously, no single observation ever suffices to deduce \( T_k(\succeq) \), neither in the constructed procedure nor in any other one, hence we can conclude that \( n = 2 \).

Case 2: \( k \) odd and \( k < m_X/2 \). The construction is the same as for case 1, except that \( f_2 = \{ a_1, \ldots, a_{(k-1)/2}, b_{(k+1)/2+1}, \ldots, b_k, b_{k+1} \} \), where \( b_{k+1} \) exists because \( k < m_X/2 \). The arguments about deducing \( T_k(\succeq) \) are also the same, starting with a comparison of \( a_1 \) and \( b_{k-1} \), except that the iteration might arrive at \( a_{(k-1)/2} \succeq b_{(k+1)/2+1} \), in which case \( T_k(\succeq) \) consists of \( a_1, \ldots, a_{(k-1)/2} \) and those \( (k+1)/2 \) alternatives that \( \succeq_2 \) ranks top within \( X \setminus f_2 \).
Case 3: $k$ odd and $k = m_X/2$. The construction is the same as for case 1, except that $f_2 = \{a_1, \ldots, a_{(k+1)/2}, b_{(k+1)/2+1}, \ldots, b_k\}$. The arguments about deducing $T_k(\succeq)$ are also the same, starting with a comparison of $a_1$ and $b_k$, except that the iteration might arrive at $a_{(k-1)/2} \succeq b_{(k+1)/2+1}$. In this case, we can conclude that $T_k(\succeq)$ consists of $a_1, \ldots, a_{(k-1)/2}$, plus either $a_{(k-1)/2}$ or $b_{(k+1)/2}$ but never both, and those $(k-1)/2$ alternatives that $\succeq_2$ ranks top among the remaining ones in $X \setminus f_2$. Hence there exist welfare preferences $\succeq$ for which $e$ does not identify $T_k(\succeq)$ after two steps. Since the missing preference between $a_{(k+1)/2}$ and $b_{(k+1)/2}$ can be learned by having $e(\Lambda_2) = f_3$ satisfy $\{a_{(k+1)/2}, b_{(k+1)/2}\} \subseteq f_3$, we know that $n \leq 3$.

It remains to be shown that $n > 2$. Fix an arbitrary elicitation procedure $e$ and denote $e(\emptyset) = f_1 = \{a_1, \ldots, a_k\}$ and $X \setminus f_1 = \{b_1, \ldots, b_k\}$, where the numbering of the alternatives is arbitrary but fixed (remember that $k = m_X/2$). Let $\succeq_1$ be the preference given (in ranking notation) by $a_1 \ldots a_k b_1 \ldots b_k$, and consider the data set $\Lambda_1 = \{(\succeq_1, f_1)\}$ and the subsequent frame $e(\Lambda_1) = f_2$. Since $k$ is odd, it follows that at least one of the pairs $\{a_1, b_k\}$, $\{a_2, b_{k-1}\}$, $\ldots$, $\{a_k, b_1\}$ must be separated on different pages by $f_2$, i.e., there exists $l = 1, \ldots, k$ such that $a_l \in f_2$ and $b_{k-l+1} \in X \setminus f_2$ or vice versa. Depending on the value of $l$, we now construct two welfare preferences $\succeq'$ and $\succeq''$. If $l = 1$, let

$$\succeq': b_1 \ldots b_{k-1} b_k a_1 a_2 \ldots a_k,$$

$$\succeq'': b_1 \ldots b_{k-1} a_1 b_k a_2 \ldots a_k.$$ 

If $l = 2, \ldots, k-1$, let

$$\succeq': a_1 \ldots a_{l-1} b_1 \ldots b_{k-l} b_{k-l+1} a_l a_{l+1} \ldots a_k b_{k-l+2} \ldots b_k,$$

$$\succeq'': a_1 \ldots a_{l-1} b_1 \ldots b_{k-l} a_l b_{k-l+1} a_{l+1} \ldots a_k b_{k-l+2} \ldots b_k.$$ 

If $l = k$, let

$$\succeq': a_1 \ldots a_{k-1} b_1 a_k b_2 \ldots b_k,$$

$$\succeq'': a_1 \ldots a_{k-1} a_k b_1 b_2 \ldots b_k.$$ 

For the two constructed welfare preferences $\succeq'$ and $\succeq''$, the elicitation procedure first generates the above described data set $\Lambda_1$. Subsequently, it generates the same data set $\Lambda_2 = \{(\succeq_1, f_1), (\succeq_2, f_2)\}$, because $\succeq'$ and $\succeq''$ differ only with respect to $a_l$ and $b_{k-l+1}$, which is not revealed by frame $f_2$. Since $T_k(\succeq') \neq T_k(\succeq'')$, it follows that $n(e, \succeq') > 2$, which implies $\max_{\succeq_\in P} n(e, \succeq) > 2$. Since $e$ was arbitrary, it follows that $n > 2$. ■

The nudging complexity is surprisingly small for the limited search model. In particular, the complexity is not growing in the number of alternatives. This begs the question to what extent the limited search model is representative for more general models. It
obviously always holds that $n \leq m_F$ if all welfare preferences are identifiable, but the number of frames $m_F$ can be extremely large (see footnote 13 in the main text). We therefore derive a tighter bound on $n$ next. The result will rest on the insight that there is always an elicitation procedure that guarantees a reduction of the set of possible welfare preferences at each step. Since there are $m_X!$ different welfare preferences that the agent might have ex ante, an elicitation procedure that reduces the set of possible preferences at each step guarantees identification of the preference and the optimal nudge after at most $m_X! - 1$ steps. It turns out that this bound is tight, because there are models for which it is reached.

**Proposition 9** Any behavioral model with identifiable preferences satisfies $n \leq m_X! - 1$, and there exist models with $n = m_X! - 1$.

**Proof.** The result follows immediately if $m_X = 2$. Hence we fix a set $X$ with $m_X \geq 3$ throughout the proof. We denote $m = m_X!$ for convenience. 

**Step 1.** We first derive the upper bound $m_X! - 1$. Consider an arbitrary behavioral model, given by $F$ and $d$, with $m_F \geq m$ and identifiable preferences. Define 

$$ \hat{n}(e, \succeq) = \min \{ t \mid P(\Lambda_t(e, \succeq)) = \{ \succeq \} \} $$

as the first step at which procedure $e$ identifies $\succeq$, and let 

$$ \hat{n} = \min \max_{e \in \Pi} \hat{n}(e, \succeq). $$

It follows immediately that $n \leq \hat{n}$, because $P(\Lambda_t(e, \succeq)) = \{ \succeq \}$ implies $G(\Lambda_t(e, \succeq)) \neq \emptyset$. We will establish the inequality $\hat{n} < m$.

Consider any $e$ and suppose $\hat{n}(e, \succeq) \geq m$ for some $\succeq \in P$. Since $|P| = m$, there must exist $k \in \{0, 1, \ldots, m - 2\}$ such that

$$ P(\Lambda_k(e, \succeq)) = P(\Lambda_{k+1}(e, \succeq)). $$

Denoting $e(\Lambda_k(e, \succeq)) = \tilde{f}$ and $d(\succeq, \tilde{f}) = \tilde{\succeq}$, we thus have $\Lambda_{k+1}(e, \succeq) = \Lambda_k(e, \succeq) \cup \{ (\tilde{\succeq}, \tilde{f}) \}$ and $d(\tilde{\succeq}, \tilde{f}) = \tilde{\succeq}$ for all $\tilde{\succeq} \in P(\Lambda_k(e, \succeq))$. We now define elicitation procedure $e'$ by letting $e'(\Lambda) = e(\Lambda)$, except for data sets $\Lambda \in L$ that satisfy both $\Lambda_k(e, \succeq) \subseteq \Lambda$ and $f \neq \tilde{f}$ for all $(\succeq, f) \in \Lambda$, which includes $\Lambda = \Lambda_k(e, \succeq)$. For those data sets, we define 

$$ e'(\Lambda) = \begin{cases} e(\Lambda \cup \{ (\tilde{\succeq}, \tilde{f}) \}) & \text{if } |\Lambda| \leq m_F - 2, \\ \tilde{f} & \text{if } |\Lambda| = m_F - 1. \end{cases} $$

Note that $e'$ is a well-defined elicitation procedure. First, $\Lambda \cup \{ (\tilde{\succeq}, \tilde{f}) \} \in L$ holds whenever the first case applies, because $\emptyset \neq P(\Lambda) \subseteq P(\Lambda_k(e, \succeq))$ and $\Lambda$ does not yet contain an
observation of $\tilde{f}$. Second, the first case then applies repeatedly because $e(\Lambda \cup \{(\tilde{\varepsilon}, \tilde{f})\}) \neq \tilde{f}$, so that $e'$ only dictates yet unobserved frames.

Consider any $\succeq' \notin P(\Lambda_k(e, \succeq))$, so that $(\succeq_1, f) \in \Lambda_k(e, \succeq)$ and $(\succeq_2, f) \in \Lambda_k(e, \succeq)$ with $\succeq_1 \neq \succeq_2$ for some $f$. From $\Lambda_k(e, \succeq') \subseteq \Lambda_t(e, \succeq')$ and thus $\Lambda_k(e, \succeq) \not\subseteq \Lambda_t(e, \succeq')$ for all $t \geq k$, it follows that preference $\succeq'$ is unaffected by the modification of the procedure, i.e., $\Lambda_t(e', \succeq') = \Lambda_t(e, \succeq')$ for all $t \in \{0, 1, \ldots, m_F\}$, so that $\hat{\nu}(e', \succeq') = \hat{\nu}(e, \succeq')$. Now consider any $\succeq' \in P(\Lambda_k(e, \succeq))$, including $\succeq' = \succeq$. Then $\Lambda_t(e, \succeq) = \Lambda_t(e, \succeq') = \Lambda_t(e', \succeq')$ holds for all $t \leq k$. For $k < t \leq m_F - 1$, the definition of $e'$ implies that $\Lambda_t(e', \succeq')$ does not contain an observation of $\tilde{f}$, and that

$$\Lambda_t(e', \succeq') \cup \{(\tilde{\varepsilon}, \tilde{f})\} = \Lambda_{t+1}(e, \succeq').$$

Thus

$$P(\Lambda_t(e', \succeq')) = P(\Lambda_t(e', \succeq') \cup \{(\tilde{\varepsilon}, \tilde{f})\}) = P(\Lambda_{t+1}(e, \succeq')),$$

so that $\hat{\nu}(e', \succeq') = \hat{\nu}(e, \succeq') - 1$. Repeated application of this construction allows us to arrive at an elicitation procedure $e^*$ for which $\hat{\nu}(e^*, \succeq) < m$ for all $\succeq \in P$, which implies that $\hat{\nu} < m$.

**Step 2.** We now show that there exist behavioral models with $n = m_X! - 1$, by giving an example of such a model. For any $\succeq \in P$, let $o(\succeq)$ denote the opposite order of $\succeq$. We write $P = \{\succeq_1, \succeq_2, \ldots, \succeq_m\}$, where the numbering of the preferences is arbitrary but fixed. We let $F = P$ and number the frames such that $f_i = o(\succeq_i)$. Finally, let $b : P \to P$ be a bijective mapping such that $b(\succeq) \notin \{\succeq, o(\succeq)\}$, for all $\succeq \in P$. Then the distortion function is given by

$$d(\succeq, f) = \begin{cases} f & \text{if } f \neq o(\succeq), \\ b(\succeq) & \text{if } f = o(\succeq). \end{cases}$$

Note that each frame $f_i$ is non-distorting for a single preference only, the one with which it coincides. This implies $n(e, \succeq) = \hat{\nu}(e, \succeq)$ for all $e \in \Pi$ and $\succeq \in P$, and thus $n = \hat{\nu}$. We will establish the equality $\hat{\nu} = m - 1$.

Consider an arbitrary $e$. Define $i_1$ such that $e(\emptyset) = f_{i_1}$, and $i_k$ for $k = 2, 3, \ldots, m$ recursively such that $e(\Lambda_{k-1}) = f_{i_k}$ for the data set

$$\Lambda_{k-1} = \bigcup_{j=1}^{k-1} \{(f_{i_j}, f_{i_j})\}.$$ 

If $\succeq_{i_m}$ is the welfare preference, then the procedure $e$ will generate the sequence of data sets $\Lambda_t(e, \succeq_{i_m}) = \Lambda_t$ for all $t \in \{0, 1, \ldots, m - 1\}$, with $\Lambda_0 = \emptyset$. It follows from the
definition of $d$ that $P(A_t) = \{\succeq_{i_{t+1}}, \succeq_{i_{t+2}}, \ldots, \succeq_{i_m}\}$ holds for each $t \in \{0, 1, \ldots, m - 1\}$. This implies $\hat{n}(e, \succeq_{i_m}) = m - 1$, and hence $\max_{\succeq \in P} \hat{n}(e, \succeq) \geq m - 1$. Since $e$ was arbitrary, it follows that $\hat{n} \geq m - 1$. Together with the result $\hat{n} < m$ established in step 1 of the proof, this implies $\hat{n} = m - 1$. ■

The tight bound on $n$ established in Proposition 9 grows more than exponentially in the number of alternatives. This shows that nudging may quickly become infeasible despite the general identifiability of preferences.