Informational Requirements of Nudging[‡]

Jean-Michel Benkert[*] and Nick Netzer[**]

This version: April 2015
First version: November 2014

Abstract

A nudge is a non-coercive paternalistic intervention that attempts to improve choices by manipulating the framing of a decision problem. As any paternalism, it faces the difficulty of determining the appropriate welfare criterion. We propose a welfare-theoretic foundation for nudging similar in spirit to the classic revealed preference approach, by investigating a model where preferences and mistakes of an agent have to be elicited from her choices under different frames. We provide characterizations of the classes of behavioral models for which nudging is possible or impossible. For the case where nudging is possible in principle, we derive results on the required quantity of information.

Keywords: nudge, framing, behavioral welfare economics

JEL Classification: D03, D04, D60, D82

[‡]We are grateful for very helpful comments by Samuel Haefner, Igor Letina, Georg Noeldeke, Yuval Salant, Armin Schmutzler, Ron Siegel, seminar audiences at HECER Helsinki, NYU Abu Dhabi, the Universities of Basel, Surrey, and Zurich, and participants at the CESifo Area Conference on Behavioural Economics 2014. 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 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, which makes it more easily acceptable than conventional paternalistic measures. Among the best-known examples already dicussed in Thaler and Sunstein (2003) is retirement saving in 401(k) savings plans, which can be encouraged tremendously by setting the default to enrollment. Under the premise that some employees make the mistake of saving too little, such automatic enrollment with the option to opt out is welfare improving. 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 mere framing, without imposing any constraints, has made the concept of nudging also politically attractive. In the UK, for instance, the Behavioural Insights Team (or "Nudge Unit") was established in 2010 as a policy unit with the goal to develop nudge-based policies.[1]

Occasionally different people might agree what it means to make the parties in question better off. In general, however, nudging shares with any paternalism the difficulty of determining the appropriate welfare criterion (see e.g. Grüne-Yanoff, 2012). What does it mean that a frame improves choice? Is it truly in the employee's own best interest to save more or to eat more healthily? Typically, the existing literature takes criteria such as increased savings or improved health for granted, or it entirely dismisses the idea of nudging based on the welfare problem (see Goldin, 2015, for a careful and critical discussion of previous approaches). In this paper, we study a model where the welfare preference of an agent first has to be inferred from her possibly distorted choices under different frames, before the success of a nudge can be evaluated. We thus attempt to develop a welfare-theoretic foundation for nudging similar in spirit to the classic revealed preference approach. The twist is that, once we accept that "in certain contexts, people are prone to error" (Sunstein, 2014, p. 4), we should be able to learn about these errors, because choices can reveal both preferences and mistakes.[2] Our basic model is a variant of Rubinstein and Salant (2012) (henceforth RS). They formulate a general framework for eliciting an agent's welfare preference from choices that are generated by a decision making process

[1] See http://www.behaviouralinsights.co.uk. The Behavioural Insights Team was privatized in 2014.

[2] Kőszegi and Rabin (2008b) first emphasized the possibility of recovering both welfare preferences and implementation mistakes from choice data, for a given behavioral theory. 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. Caplin and Martin (2012) provide conditions under which welfare preferences can be recovered from choice data in a setting where frames contain payoff-relevant information, such that framing effects are fully rational.

with mistakes, which can be affected by frames. RS investigate the problem of learning about the welfare preference from a data set that contains observations of behavior and, possibly, frames. We follow their approach in a first step, and then proceed to evaluating the frames based on the acquired knowledge about the agent's welfare preference.

In the model there is an agent with an unobservable (strict) *welfare preference* $\succ$ that represents the normatively relevant well-being of the agent. The decision making process, or behavioral model, is summarized by a distortion function d, which yields a *behavioral preference* $d(\succ, f)$ for each possible combination of a welfare preference $\succ$ and a frame f. The interpretation is that the agent acts as if maximizing $d(\succ, f)$ instead of $\succ$ if the decision situation is framed by f. To illustrate, consider an agent with welfare preference $c \succ a \succ b \succ d$ over the set of alternatives $X = \{a, b, c, d\}$. Let the agent behave according to the model of *perfect recall satisficing* as in RS. She is satisfied with any of the $k = 2$ top ranked alternatives; in this case, c and a are satisfactory. The frame f describes the order in which the alternatives are presented. When choosing from a non-empty subset $S \subseteq X$ (e.g. the budget set), the agent considers the alternatives in S sequentially and picks whichever satisfactory alternative is presented first. If S turns out not to contain any satisfactory alternative, the agent reconsiders all alternatives in S and chooses according to her welfare preference. Suppose presentation is in alphabetical order. Because a is presented before c, the agent will choose a whenever $a \in S$, even if $c \in S$, 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. Taken together, these choices look as if the agent was maximizing the preference $a \succ c \succ b \succ d$. Suppose this behavioral preference is observed in the standard revealed preference sense, by observing the agent's choices from different subsets $S \subseteq X$ under the fixed frame of alphabetical presentation. Reversing the distortion process then allows us to conclude that the agent's welfare preference must be either $a \succ c \succ b \succ d$ or $c \succ a \succ b \succ d$; these two but no other welfare preferences generate the observed behavior for the given decision making process.

Based on this knowledge let us turn to the problem of nudging, which here amounts to determining the optimal order of presentation. Any order that presents a before c would be optimal if the agent's welfare preference was $a \succ c \succ b \succ d$, but induces the above described decision mistake between a and c if the welfare preference is $c \succ a \succ b \succ d$. 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. Unfortunately, the problem cannot be solved by observing the agent under additional frames. The order of presentation fully determines the agent's 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 frame, nudging runs into irresolvable information problems.

But now consider an alternative decision making process, a model of *limited search*.

When the agent looks for a product online, all alternatives $X = \{a, b, c, d\}$ are displayed by a search engine, two of them on the first result page and the other two on the second page. The frame f here is the set of alternatives presented on the first page. As before, let $S \subseteq X$ denote the subset from which the agent can actually choose, e.g. those alternatives among X that are affordable or in stock. Whenever the first result page contains at least one of these available alternatives, then the agent does not even look at the second page but chooses from $S \cap f$ according to her welfare preference. Only if none of the elements of S are displayed on the first page, the agent moves on to the second page and chooses according to her welfare preference there. Suppose the welfare preference is $c \succ a \succ b \succ d$ as in the previous example, and let the first page be given by $f = \{a, b\}$. Then this agent will choose a whenever $a \in S$ even if $c \in S$, because c is displayed only on the second page. 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 $a \succ b \succ c \succ d$. Suppose again that this behavioral preference is revealed (by observation of the agent's choices from different subsets $S \subseteq X$, given the fixed frame). Reversing the distortion process now unveils that the agent must truly prefer a over b and c over d, which leaves us with the six possible welfare preferences marked in the first column of Table 1.

	$\{a, b\}$: $a \succ b \succ c \succ d$	$\{a, d\}$: $a \succ d \succ c \succ b$
$a \succ b \succ c \succ d$	✓	
$a \succ c \succ b \succ d$	✓	✓
$a \succ c \succ d \succ b$	✓	✓
$c \succ a \succ b \succ d$	✓	✓
$c \succ a \succ d \succ b$	✓	✓
$c \succ d \succ a \succ b$	✓	
$a \succ d \succ c \succ b$		✓
$c \succ b \succ a \succ d$		✓

Table 1: Reversing Limited Search

A nudge should place the two welfare-best alternatives on the first page, thus helping the agent avoid decision mistakes like the ones between a or b and c above. Unfortunately, each of the four alternatives belongs to the top two for at least one possible welfare preference, but none of them for all possible welfare preferences. Hence no frame guarantees fewer mistakes than any other. In contrast to the satisficing example, however, gathering more information helps. Observing the agent's choices under frame $\{a, d\}$ reveals the behavioral preference $a \succ d \succ c \succ b$, from which the six welfare candidates marked in the second column of Table 1 can be deduced. The four welfare preferences that are consistent with the observations from both frames now all agree that a and c are the two best alternatives. Hence we know that $\{a, c\}$ is the optimal nudge. 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.

The two examples illustrate that the scope of nudging depends on the behavioral model d, which we proceed to examine in our general framework (Section 2). Given a behavioral model and a data set containing observations of frames and revealed behavioral preferences, we first perform the reverse learning procedure to narrow down the set of possible welfare preferences. We then compare frames pairwise, saying that f is a (weakly) successful nudge over f' if all choices under f are at least as good as under f', for all of the remaining possible welfare preferences. A frame is optimal if it is a (weakly) successful nudge over all the other frames. As a first result, we show in Section 3 that the ability to identify an optimal frame coincides with the ability to identify the welfare preference: an optimal frame is revealed by some sufficiently rich data set if and only if the welfare preference is revealed by some sufficiently rich data set. This result does not say that the welfare preference actually has to be learned for successful nudging, as the previous example has shown, but it allows us to consider two polar cases: models where the welfare preference can never be identified, as in the satisficing example, and models where the welfare preference can be identified, as in the limited search example.

If the welfare preference cannot be identified, then finding an optimal frame is out of reach. In Section 4 we pursue the more modest goal of identifying frames which are dominated by others. Such dominated frames can exist, as we will show by example. However, if the behavioral model satisfies a property that we term the *frame cancellation property*, then all frames are always undominated, irrespective of the data set's richness. With the frame cancellation property, we can never learn from observing framed choices what we would need to know to improve these choices. Several important models have the frame cancellation property. A first example is perfect recall satisficing in its general formulation. A second example is the much-discussed case where the agent chooses the one alternative out of two that is marked as a default, as for the 401(k) savings plans. We also discuss models of choice from lists (Rubinstein and Salant, 2006) that have the frame cancellation property. Finally, we present a decision making procedure with limited sensitivity that nests all these behavioral models.

If, by contrast, the welfare preference can ultimately be learned, then questions of complexity arise. How many, and which, observations are necessary to determine the optimal frame? In Section 5 we define an *elicitation procedure* as a rule that specifies the next frame under which we want to observe the agent, for each history of observations. Holding fixed the welfare preference of the agent, an elicitation procedure generates a sequence of expanding data sets that eventually identifies the optimal frame. We define the complexity n of the nudging problem as the minimum over all elicitation procedures of the number of observations after which the optimal frame is guaranteed to be known. As a first application, we construct an optimal elicitation procedure for the limited search model in its general formulation. We show that $n = 2$ or $n = 3$, depending on the

number of alternatives and the capacity of the search engine's result pages. Thus learning and nudging are relatively simple in this specific model. As a general result, we then establish a bound on n for arbitrary behavioral models. The bound, which is reached by some models, corresponds to the number of possible welfare preferences and thus grows more than exponentially in the number of alternatives. This implies that the informational requirements of nudging can easily become prohibitive even with identifiable welfare preferences.

Several of our results reveal informational limitations for anyone attempting to base the selection of nudges on a solid welfare-theoretic foundation. The identification of an optimal nudge is often impossible, and it is an observationally heavy task in other cases. This is even more so the case, as some aspects of our model work in favor of nudging. For instance, the assumption that the behavioral model is known to the regulator makes it particularly easy to learn about welfare (see Section 6 for an extension to model uncertainty and for a theory-free approach). The same applies to the assumption that each combination of a welfare preference and a frame generates a unique behavioral preference, and that this preference together with the frame is perfectly observable (see Section 6 for an extension to imperfectly observable frames). At the same time, our analysis reveals that seemingly minor differences between behavioral models – such as whether an agent's failure to optimize is due to a low aspiration level or due to a restricted number of considered alternatives – can have profoundly different consequences for the ability to improve well-being by framing. This points at important questions for future research on decision procedures.

Goldin and Reck (2015) also study the problem of identifying welfare preferences when choices are distorted by frames. They focus mostly on binary choice problems with two frames (e.g. defaults) and aim at estimating the population shares of consistent and inconsistent agents and the two preference types. The share of consistent agents who prefer a given alternative is equal to the share of agents who choose this alternative "against the frame" (p. 12) in a representative sample. The preference shares among the inconsistent agents can then be deduced from this information under certain identifying assumptions, for instance when they are identical to the consistent agents (after controlling for observable differences) or when there is additional information about the preferences of the entire population. If these assumptions are valid, it is possible to identify the frame that induces the best choice for a majority of the population (see also Goldin, 2015). Such informational requirements are not the only obstacle that a libertarian paternalist has to overcome. Spiegler (2015), for instance, 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 (e.g. the definition of a default product).

2 Model

The framework is a variant of RS. Let X be a finite set of welfare-relevant 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 $\succeq \in P$. 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 considered welfare-irrelevant.[3] The agent's behavior is summarized by a distortion function $d : P \times F \to P$, which assigns a distorted order $d(\succeq, f) \in P$ to each combination of $\succeq \in P$ and $f \in F$. The interpretation is that an agent with true welfare preference $\succeq$ acts as if maximizing the preference $d(\succeq, f)$ if the decision situation is framed by f.[4] The distortion function represents a conjecture about the relation between welfare, frames and choice. Such a conjecture typically relies on insights about the decision-making process and thus originates from non-choice data, which is becoming increasingly more available (e.g. also from neuroscience or psychology). For instance, eye-tracking or monitoring of browsing behaviors can provide the type of information necessary to substantiate models like our limited search example. Arguably, non-choice-based conjectures about the relation between choice and welfare always have to be invoked, even in standard welfare economics.[5] Before we proceed, we formally present the two behavioral models that were used in the introductory examples.

Model 1 (Perfect Recall Satisficing) *Alternatives are presented sequentially and the frame $f \in F = P$ determines their order. From any non-empty subset $S \subseteq X$ the agent chooses the first alternative that exceeds her aspiration level $k \in \{2, \dots, m_X\}$, i.e., that is among the top k alternatives according to her welfare preference. If no element of S turns out to exceed this threshold, then the agent chooses the welfare-optimal one. Choices between satisfactory alternatives will thus always be in line with the order of presentation, and all other choices are in line with the welfare preference. Hence we can obtain $d(\succeq, f)$ directly from $\succeq$ by rearranging the top k elements according to their order in f. In 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.*

[3] For specific applications, the modeller has to judge which dimensions are welfare-relevant and which are not. For instance, it appears 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 information about the decision problem.

[4] This assumes that, given any frame, choices are consistent and can be represented by a preference. Salant and Rubinstein (2008) refer to extended choice functions with this property as "salient consideration functions" (p. 1291). The assumption rules out behavioral models in which choices violate standard axioms already when a frame is fixed. De Clippel and Rozen (2014) investigate the problem of learning from incomplete data sets without such an assumption.

[5] See Kőszegi and Rabin (2007, 2008a) and Rubinstein and Salant (2008). For an opposing perspective and a critical discussion of the ability to identify the decision process, see Bernheim (2009).

Model 2 (Limited Search) *All alternatives in X are displayed by a search engine, on either the first or the second result page. The frame f 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. When the agent chooses from a non-empty subset $S \subseteq X$ (not all displayed alternatives might be in stock or affordable), she remains on the first page whenever $S \cap f$ is non-empty, i.e., when the first page contains at least one available alternative, and she chooses according to her welfare preference. If none of the first page alternatives belongs to S, she moves to the second page and chooses optimally there. Choices between alternatives on the same page will thus always be in line with the welfare preference, but any alternative on the first page is revealed preferred 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. This model is similar to the gradual accessibility model in Salant and Rubinstein (2008), but the eventual choice rule is different.*

The only assumption that we impose on the behavioral model in general is that for each $\succeq \in P$ there exists $f \in F$ such that $d(\succeq, f) = \succeq$. This rules out that some preferences are distorted by all possible frames and allows us to focus on the informational requirements of nudging, without having to deal with exogenously unavoidable distortions. The assumption does not imply the existence of a neutral frame that is non-distorting for all preferences.[6] 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.

A behavioral data set is a subset $\Lambda \subseteq P \times F$. The interpretation is that we observe behavioral preferences in the usual revealed preference sense (as a result of observing choices from sufficiently many different subsets $S \subseteq X$ to recover the preference), and possibly we do this for several frames.[7] Further following RS, we say that $\succeq$ is consistent with Λ if for each $(\succeq', f') \in \Lambda$ it holds that $\succeq' = d(\succeq, f')$. In that case, $\succeq$ is a possible welfare preference, because the data set might have been generated by an agent with that

[6] 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. The existence of such a frame makes the welfare elicitation problem and also the nudging problem straightforward. Often, however, this solution is not available, e.g. defaults are unavoidable for organ donations, and alternatives must always be presented in some order or arrangement.

[7] Formally, 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. See Section 6 for the respective generalization.

preference. Let

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

be the maximal data set that can be observed if the welfare preference is $\succeq$. Then the set of all welfare preferences that are consistent with Λ is given by

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

Without further mention, we consider only data sets Λ 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.[8] Observe that a frame f cannot appear more than once in such data sets. Observe also that $P(\varnothing) = P$ 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 Λ. Once the set of possible welfare preferences is narrowed down to $P(\Lambda)$, previously different frames might have become behaviorally equivalent. Thus, for any f let

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

be the equivalence class of frames for f, i.e., 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. We now compare the elements of $F(\Lambda)$ from the perspective of the possible welfare preferences, based on the choices that they induce. For any $\succeq$ and any non-empty $S \subseteq X$, let $c(\succeq, S)$ be the element of S that is chosen by an agent who maximizes $\succeq$.

Definition 1 *For any f, f' and Λ, $[f]_\Lambda$ is a weakly successful nudge over $[f']_\Lambda$, 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 f' (and all equivalent ones) no matter which of the remaining welfare preferences is the true one. The binary nudging relation $N(\Lambda)$ shares

[8] RS derive conditions under which data sets do or do not falsify a model conjecture. A falsified model is of no use for the purpose of nudging and would have to be replaced by a conjecture for which $P(\Lambda)$ is non-empty.

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). A major difference is that the multiplicity of preferences here reflects a lack of information, not multiple selves. Thus, 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. In fact, the only way in which the data set Λ matters for the nudging relation is via the set $P(\Lambda)$.

The following Lemma 1 summarizes additional properties of $N(\Lambda)$ that will be useful. 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.[9] For instance, in the satisficing example from the introduction, where the welfare preference was given by $c \succ a \succ b \succ d$ and alphabetical order of presentation resulted in the behavioral preference $a \succ c \succ b \succ d$, we would obtain $B(\succeq, f) = \{(a, c)\}$. For the limited search example where frame $\{a, b\}$ distorted the same welfare preference to $a \succ b \succ c \succ d$, we would obtain $B(\succeq, f) = \{(a, c), (b, c)\}$.

Lemma 1 *(i)* $[f]_\Lambda N(\Lambda)[f']_\Lambda$ *if and only if* $B(\succeq, f) \subseteq B(\succeq, f')$ *for each* $\succeq \in P(\Lambda)$. *(ii)* $N(\Lambda)$ *is a partial order (reflexive, transitive, antisymmetric) on* $F(\Lambda)$.

Proof. (i) Suppose that $B(\succeq, f) \subseteq B(\succeq, f')$ holds for each $\succeq \in P(\Lambda)$. To show that $[f]_\Lambda N(\Lambda)[f']_\Lambda$, 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 $(x, y) \notin \succeq$, hence $(x, y) \notin d(\succeq, f')$. Then $c(d(\succeq, f'), \{x, y\}) = y \succeq x = c(d(\succeq, f), \{x, y\})$, which implies that $[f]_\Lambda N(\Lambda)[f']_\Lambda$ does not hold, by Definition 1.

(ii) Reflexivity and transitivity of $N(\Lambda)$ follow from the set inclusion characterization in statement (i). To show antisymmetry, consider any $f, f' \in F$ with $[f]_\Lambda N(\Lambda)[f']_\Lambda$ and $[f']_\Lambda N(\Lambda)[f]_\Lambda$. By (i) this is equivalent to $B(\succeq, f) = B(\succeq, f')$ and thus $d(\succeq, f) = d(\succeq, f')$ for each $\succeq \in P(\Lambda)$, hence $[f]_\Lambda = [f']_\Lambda$. ■

Since $B(\succeq, f)$ describes all the mistakes in binary choice that frame f causes for welfare preference $\succeq$, statement (i) of the lemma formalizes the intuition that a successful nudge is a frame that induces fewer mistakes. Statement (ii) implies that the binary relation is sufficiently well-behaved to consider different notions of optimality.

[9] Even though we often represent preferences as rankings like $c \succ a \succ b \succ d$, we remind ourselves that technically both $d(\succeq, f)$ and $\succeq$ are subsets of the set of ordered pairs $X \times X$.

3 Nudgeability

An optimal nudge is a frame that guarantees (weakly) better choices than all the other frames. Let

$$G(\Lambda) = \{f \mid [f]_\Lambda N(\Lambda)[f']_\Lambda, \forall f' \in F\}$$

be the set of frames which have been identified as optimal by the data set Λ. Formally, $G(\Lambda)$ coincides with the greatest element of the partially ordered set $F(\Lambda)$, and it might be empty. Since the nudging relation becomes more complete as we collect additional observations, it follows that optimal frames are more likely to exist for larger data sets. Therefore, the following result provides a necessary and sufficient condition for the existence of an optimal frame based on the maximal data set. The result is relatively straightforward, but important as it will allow us to classify behavioral models according to whether the search for an optimal frame is promising or hopeless.

Definition 2 *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.*

Proof. 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 $B(\succeq, f) = \varnothing \subseteq B(\succeq, f')$ and hence $[f]_{\bar{\Lambda}(\succeq)} N(\bar{\Lambda}(\succeq))[f']_{\bar{\Lambda}(\succeq)}$ by Lemma 1, which implies $f \in G(\bar{\Lambda}(\succeq))$. For the converse, 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 $B(\succeq, f_1) = \varnothing$ and $B(\succeq', f_2) = \varnothing$. Assume by contradiction that there exists $f \in G(\bar{\Lambda}(\succeq))$. Then $[f]_{\bar{\Lambda}(\succeq)} N(\bar{\Lambda}(\succeq))[f_1]_{\bar{\Lambda}(\succeq)}$ must hold, which implies $B(\succeq, f) = \varnothing$ by Lemma 1, and hence $d(\succeq, f) = \succeq$. The analogous argument for f_2 implies $d(\succeq', f) = \succeq'$, which contradicts that $\bar{\Lambda}(\succeq') = \bar{\Lambda}(\succeq)$, i.e., that $\succeq$ is not identifiable. ■

The if-statement is immediate: an identifiable welfare preference is known for sure once the maximal data set has been collected, and all the non-distorting frames are optimal with that knowledge. It is worth emphasizing again, however, that the result does not imply that the welfare preference actually has to be learned perfectly for successful nudging. It only tells us that, if $\succeq$ is the true and identifiable welfare preference, then for some sufficiently large data set Λ we will be able to identify an optimal nudge; the set $P(\Lambda)$ of consistent welfare preferences might still contain more than one element at that point. The only-if-statement tells us that there is no hope to ever identify an optimal frame if the welfare preference cannot be identified, i.e., if there exists another welfare preference

$\succeq'$ that is behaviorally equivalent to $\succeq$ under all frames. In this case we say that $\succeq$ and $\succeq'$ are indistinguishable. A frame could then only be optimal if it does not distort any of the two, but this is impossible as such a frame would generate different observations for $\succeq$ and $\succeq'$ and hence would empirically discriminate between them.

In the following, we consider the two polar cases of behavioral models where all welfare preferences are identifiable or non-identifiable, respectively. Our prime example for non-identifiable preferences is the perfect recall satisficing model. Any two 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. We know that the nudging relation will never admit an optimal frame in that case, but we might still be able to exclude some frames that are dominated by others. Our prime example for identifiable preferences is the limited search model (if $m_X \geq 3$). There, we learn the welfare preference among all alternatives on the same page, and thus we can identify the complete welfare preference by observing behavior under sufficiently many different frames. In that case, we will be interested in the required quantity of information and the properties of optimal learning procedures.

4 Non-Identifiable Preferences

Our previous notion of optimality is strong, as it requires an optimal frame to outperform all others. If such a frame does not exist, we can weaken optimality to the requirement that a reasonable frame should not be dominated by another one. Let

$$M(\Lambda) = \{f \mid [f']_\Lambda N(\Lambda)[f]_\Lambda \text{ only if } f' \in [f]_\Lambda\}$$

be the (always non-empty) set of frames which are undominated, based on our knowledge of the data set Λ. Formally, $M(\Lambda)$ is the union of all elements that are maximal in the partially ordered set $F(\Lambda)$. A frame which is not in $M(\Lambda)$ can be safely excluded, as there exists a nudge that guarantees an improvement over it.

Dominated frames can exist already ex ante with no knowledge of the agent's welfare preference. For instance, certain informational arrangements could be interpreted as being dominant over others because they objectively clarify the available information and improve the decision quality (e.g. Camerer et al., 2003). In the following example we show that ex ante undominated frames can become dominated for richer knowledge, too. Assume that $X = \{a, b, c, d\}$ and consider the distortion function for the four preferences and three frames depicted in Figure 1.[10] The two welfare preferences $\succeq_1$ and $\succeq_2$ are

[10]The example focusses on only four welfare preferences, but it can be expanded to encompass the set of all possible preferences. We can also add additional frames without changing its insight.

indistinguishable, as each frame maps them into the same distorted preference, and the same holds for $\succeq_3$ and $\succeq_4$. Note also that none of the frames is dominated before any data has been collected, $M(\varnothing) = \{f_1, f_2, f_3\}$, because each one is the unique non-distorting frame for one possible welfare preference. Now suppose we observe $\Lambda = \{(\succeq_2, f_2)\}$, so that $P(\Lambda) = \{\succeq_1, \succeq_2\}$. It follows immediately that none of the potentially non-distorting frames f_1 and f_2 is dominated. The frame f_3, however, is now dominated by f_1. If the welfare preference is $\succeq_2$, then f_1 induces a mistake between a and b, but so does f_3, which induces an additional mistake between c and d. Hence we obtain $M(\Lambda) = \{f_1, f_2\}$. We have learned enough to identify a nudge over f_3, but no additional observation will ever allow us to compare f_1 and f_2.

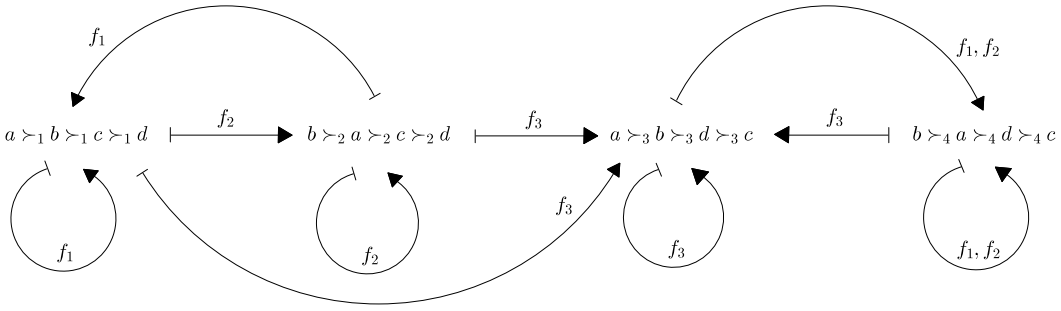

Figure 1: Dominated Frame f_3

Frame f_3 is particular, as it maps the indistinguishable set $\{\succeq_1, \succeq_2\}$ outside of itself. If a maximal set of preferences that are indistinguishable from the actual welfare preference is closed under the distortion function for some frame, then that frame cannot be dominated. This follows because the behavioral preference observed under the frame must itself be considered as a possible welfare preference, and the frame is non-distorting for it. This observation provides the basis for the following definition and result.

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. It follows that preference $d(\succeq, f)$ is observationally equivalent to $\succeq$, for any $\succeq \in P$ and $f \in F$, and hence all maximal indistinguishable sets of preferences are closed under the distortion function for any given frame.

Proposition 2 *If d has the frame-cancellation property, then $M(\Lambda) = F$ for all Λ.*

Proof. Consider any d with the frame-cancellation property and any data set Λ. Fix any frame $f_1 \in F$, and let $f_2 \in F$ be an arbitrary frame with $f_2 \notin [f_1]_\Lambda$. Then, by definition of $[f_1]_\Lambda$, 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 $\succeq_1 \in P(\Lambda)$. We also obtain $d(\succeq_1, f_1) = d(\succeq, f_1) = \succeq_1$, which implies $B(\succeq_1, f_1) = \varnothing$. From $\succeq_1 \neq \succeq_2$ and the frame-cancellation property, it follows that

$$B(\succeq_1, f_2) = d(\succeq_1, f_2) \setminus \succeq_1 = d(d(\succeq, f_1), f_2) \setminus \succeq_1 = d(\succeq, f_2) \setminus \succeq_1 = \succeq_2 \setminus \succeq_1 \neq \varnothing.$$

Hence $B(\succeq_1, f_1) \subset B(\succeq_1, f_2)$, and Lemma 1 implies that $[f_2]_\Lambda N(\Lambda)[f_1]_\Lambda$ does not hold. Since f_2 was arbitrary we conclude that $f_1 \in M(\Lambda)$, and, since f_1 was arbitrary, that $M(\Lambda) = F$. ■

If the frame-cancellation property holds, then irrespective of how many data points we have collected, we will never know enough to exclude even a single dominated frame.[11] To provide an analogy, we can think of $M(\Lambda)$ as the set of Pareto efficient policies, because moving away from any $f \in M(\Lambda)$ may make the agent better off with respect to some $\succeq \in P(\Lambda)$ only at the cost of making her worse off with respect to some other $\succeq' \in P(\Lambda)$. Proposition 2 therefore states that all frames are Pareto efficient. If we want to select between them, we may need to resort to approaches that can be used to compare Pareto efficient allocations, involving stronger assumptions such as probabilistic priors and comparable cardinal utilities. We will return to this idea in the conclusion.

Is frame-cancellation a plausible condition? Notice first that Proposition 2 does not require preferences to be non-identifiable. However, identifiability of $\succeq$ is consistent with the frame-cancellation property only if $d(\succeq, f) = \succeq$ for all $f \in F$, i.e., all frames must be equally non-distorting for $\succeq$. This corresponds to the standard rational choice model. Another 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. This 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.

[11] Formally, the binary relation $N(\Lambda)$ is always diagonal, i.e., empty except for its reflexive component.

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.[12] An example in which this holds is satisficing without recall (see also RS). 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. It is easy to verify that this model has the frame-cancellation property.

We conclude the section by presenting a class of decision processes with limited sensitivity that nests all these examples of models with the frame-cancellation property.

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 $(k_1, k_2, \ldots, k_s)$ of positive integers with $\sum_{i=1}^{s} k_i = m_X$. A welfare preference $\succeq$ induces a 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, let $P_{\succeq}$ be the set of welfare preferences that induce the same partition of X as $\succeq$, for any $\succeq \in P$. Then $d(\succeq', f) = d(\succeq'', f) \in P_{\succeq}$ must hold whenever $\succeq', \succeq'' \in P_{\succeq}$, for all $f \in F$. Any such function satisfies the frame-cancellation property.*[13] *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 $(k_1, k_2, \ldots, k_s) = (1, 1, \ldots, 1)$, 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 $k_1 = m_X$.*

[12] 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).

[13] For any $\succeq \in P$, since $\succeq \in P_{\succeq}$ holds we have $d(\succeq, f_1) \in P_{\succeq}$ 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 $P_{\succeq}$ in the limited sensitivity model, but these sets will not in general be generated by some vector $(k_1, k_2, \ldots, k_s)$ as required by the limited sensitivity model.

5 Identifiable Preferences

We now turn to the case of identifiable welfare preferences, which guarantees knowledge of an optimal nudge once a maximal data set has been observed. Collecting a maximal data set requires observing the agent under all m_F frames, however, which might be beyond our means. We are thus interested in optimal data gathering procedures. The idea is that a regulator, who ultimately seeks to impose the optimal nudge, is also able to impose a specific sequence of frames with the goal of eliciting the agent's welfare preference.

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

$$L_s = \{\Lambda | P(\Lambda) \neq \varnothing \text{ and } |\Lambda| = s\}$$

be the collection of all data sets that do not falsify the behavioral model and contain exactly s observations, i.e., observations for s different frames. In particular, $L_0 = \{\varnothing\}$, and L_{m_F} consists of all maximal data sets. 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 4 *An elicitation procedure is a mapping $e : L \to F$ 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 $e(\varnothing)$ and, if the welfare preference is $\succeq$, generates the first data set $\Lambda_1(e, \succeq) = \{(d(\succeq, e(\varnothing)), e(\varnothing))\}$. 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 $\Lambda_0(e, \succeq) = \varnothing$ and

$$\Lambda_{s+1}(e, \succeq) = \Lambda_s(e, \succeq) \cup \{(d(\succeq, e(\Lambda_s(e, \succeq))), e(\Lambda_s(e, \succeq)))\},$$

until the maximal data set $\Lambda_{m_F}(e, \succeq) = \bar{\Lambda}(\succeq)$ 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. Notice that an elicitation procedure dictates frames also for pre-collected data sets that it never generates. We tolerate this redundancy because otherwise definitions and proofs would become substantially more complicated, at no gain. Now define

$$n(e, \succeq) = \min\{s \mid G(\Lambda_s(e, \succeq)) \neq \varnothing\}$$

as 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 $\max_{\succeq \in P} n(e, \succeq)$ steps. With

E denoting the set of all elicitation procedures, we thus have to be prepared to gather

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

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.

Proposition 3 *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. See Appendix. ■

To understand our construction of an optimal elicitation procedure for the limited search model, consider again the simple example from the Introduction. The procedure starts with an arbitrary frame, $f_1 = \{a, b\}$, and generates a first behavioral preference, $a \succ_1 b \succ_1 c \succ_1 d$. We now know that the welfare preference satisfies $a \succ b$ and $c \succ d$. The second frame is constructed by taking the top element from f_1 and the bottom element from $X \backslash f_1$, which yields $f_2 = \{a, d\}$. From the induced behavioral preference $a \succ_2 d \succ_2 c \succ_2 b$ we can learn that $a \succ d$ and $c \succ b$. This information is enough to deduce that a and c are the two welfare-optimal alternatives, because both b and d are worse than each of them. If instead we had learned that $a \succ d$ and $b \succ c$ at the second step, we could have concluded that a and b are optimal. If we had learned that $d \succ a$, we could similarly have concluded that c and d are optimal.

This argument can be generalized. Starting with an arbitrary frame f_1, we learn the welfare preference within the sets f_1 and $X \backslash f_1$. Denote the elements of f_1 in descending welfare order by $a_1, a_2, \ldots, a_k$ and the elements of $X \backslash f_1$ in descending welfare order by $b_1, b_2, \ldots, b_{m_X - k}$. When $k = m_X/2$ and k is even, for instance, the second frame is constructed to contain the $k/2$ best alternatives from f_1 and the $k/2$ worst alternatives from $X \backslash f_1$, which yields $f_2 = \{a_1, \ldots, a_{k/2}, b_{k/2+1}, \ldots, b_k\}$ (this construction has to be adjusted slightly for different values of k and m_X). After having learned the welfare preference within the sets f_2 and $X \backslash f_2$, we can deduce the k welfare-best alternatives and thus the optimal nudge as follows. We consider the just learned welfare preference among a_l and b_{k-l+1} successively for $l = 1, \ldots, k/2$. Whenever $a_l \succeq b_{k-l+1}$, we can conclude that b_{k-l+1} does not belong to the optimal nudge (because the k alternatives $a_1, \ldots, a_l, b_1, \ldots, b_{k-l}$ are welfare better) while a_l does belong to it (because the k alternatives $a_{l+1}, \ldots, a_k, b_{k-l+1}, \ldots, b_k$ are welfare worse). On the first instance of l with

$b_{k-l+1} \succeq a_l$ we thus know that the optimal nudge is $\{a_1, \ldots, a_{l-1}, b_1, \ldots, b_{k-l+1}\}$. If this does not occur up to and including $l = k/2$, we know that the optimal nudge consists of $a_1, \ldots, a_{k/2}$ and the $k/2$ best alternatives from $X \backslash f_2$.

For more general behavioral models, it obviously holds that $n \leq m_F$ if all welfare preferences are identifiable, but the number of frames m_F can be arbitrarily large. We therefore derive a bound on n next. The following result rests on the insight that there is always an elicitation procedure that guarantees a reduction of the set of possible welfare preferences at each step, while there are models for which a reduction can be guaranteed by only one preference at each step.

Proposition 4 *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. To establish the inequality, consider an arbitrary behavioral model, given by F and d, with $m_F \geq m$ and identifiable preferences. Define

$$\hat{n}(e, \succeq) = \min\{s \mid P(\Lambda_s(e, \succeq)) = \{\succeq\}\}$$

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

$$\hat{n} = \min_{e \in E} \max_{\succeq \in P} \hat{n}(e, \succeq).$$

It follows immediately that $n \leq \hat{n}$, because $P(\Lambda_s(e, \succeq)) = \{\succeq\}$ implies $G(\Lambda_s(e, \succeq)) \neq \varnothing$. 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 $d(\succeq', \tilde{f}) = \tilde{\succeq}$ for all $\succeq' \in P(\Lambda_k(e, \succeq))$. 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

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

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 $\varnothing \neq P(\Lambda) \subseteq P(\Lambda_k(e, \succeq))$ and Λ does not yet contain an

observation of $\tilde{f}$. Second, the first case then applies repeatedly because $e(\Lambda \cup \{(\tilde{\succeq}, \tilde{f})\}) \neq \tilde{f}$, 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 $\Lambda_k(e, \succeq') \subseteq \Lambda_s(e, \succeq')$ and thus $\Lambda_k(e, \succeq) \nsubseteq \Lambda_s(e, \succeq')$ for all $s \geq k$, it follows that preference $\succeq'$ is unaffected by the modification of the procedure, i.e., $\Lambda_s(e', \succeq') = \Lambda_s(e, \succeq')$ for all $s \in \{0, 1, \ldots, m_F\}$, so that $\hat{n}(e', \succeq') = \hat{n}(e, \succeq')$. Now consider any $\succeq' \in P(\Lambda_k(e, \succeq))$, including $\succeq' = \succeq$. Then $\Lambda_s(e, \succeq) = \Lambda_s(e, \succeq') = \Lambda_s(e', \succeq')$ holds for all $s \leq k$. For $k < s \leq m_F - 1$, the definition of e' implies that $\Lambda_s(e', \succeq')$ does not contain an observation of $\tilde{f}$, and that

$$\Lambda_s(e', \succeq') \cup \{(\tilde{\succeq}, \tilde{f})\} = \Lambda_{s+1}(e, \succeq').$$

Thus

$$P(\Lambda_s(e', \succeq')) = P(\Lambda_s(e', \succeq') \cup \{(\tilde{\succeq}, \tilde{f})\}) = P(\Lambda_{s+1}(e, \succeq')),$$

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

Step 2. To establish the equality, we construct a model with identifiable preferences and $n = m - 1$. Write $P = \{\succeq_0, \succeq_1, \ldots, \succeq_{m-1}\}$, where the numbering of preferences is arbitrary but fixed. Let $F = \{f_i \mid i = 0, 1, \ldots, m-1\}$, so that $m_F = m$, and define d by

$$d(\succeq, f_i) = \begin{cases} \succeq_{[i+1]} & \text{if } \succeq = \succeq_i, \\ \succeq_{[i+2]} & \text{otherwise,} \end{cases}$$

where $[j]$ stands short for j mod m. Hence each frame f_i is non-distorting only for the single preference $\succeq_{[i+2]}$, which implies $n = \hat{n}$. We will establish the equality $\hat{n} = m - 1$.

Consider any e. Define i_0 such that $e(\varnothing) = f_{i_0}$, and define i_s for $s = 1, \ldots, m-1$ recursively such that $e(\Lambda_s) = f_{i_s}$ for the data set

$$\Lambda_s = \bigcup_{j=0}^{s-1} \{(\succeq_{[i_j+2]}, f_{i_j})\}.$$

It follows from the definition of d that $P(\Lambda_s) = \{\succeq_{i_s}, \succeq_{i_{s+1}}, \ldots, \succeq_{i_{m-1}}\}$ holds for each $s \in \{0, 1, \ldots, m-1\}$, where $\Lambda_0 = \varnothing$. Also, for $\succeq_{i_{m-1}}$ it holds that $\Lambda_s(e, \succeq_{i_{m-1}}) = \Lambda_s$ for all $s \in \{0, 1, \ldots, m-1\}$, which implies $\hat{n}(e, \succeq_{i_{m-1}}) = m-1$. Thus $\max_{\succeq \in P} \hat{n}(e, \succeq) \geq m-1$. Since e was arbitrary, it follows that $\hat{n} \geq m-1$. Together with the result $\hat{n} < m$ established in step 1, this implies $\hat{n} = m - 1$. ■

Since there are $m_X!$ different welfare preferences that the agent might have ex ante, an elicitation procedure that strictly 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. In the proof, we construct a behavioral model where identification of the optimal nudge actually requires identification of the preference, and this takes all $m_X! - 1$ steps. In the model, each observation of behavior under a frame either reveals a specific welfare preference to be the true one, or it excludes it from the set of possible welfare preferences. No matter in which order frames are dictated by the elicitation procedure, it is always possible that the agent's welfare preference is the one not revealed until the end. Proposition 4 is again bad news for nudging. The bound is growing more than exponentially in the number of alternatives, which may quickly make nudging infeasible despite the general identifiability of preferences.

6 Discussion and Extensions

We have made several assumptions that work in favor of nudgeability and may seem overly restrictive. Since our analysis at least sometimes reveals a simple solution to the information problem (see e.g. Proposition 3), we now relax some of these assumptions. We discuss model uncertainty (Section 6.1), a theory-free approach (Section 6.2), and imperfectly observable frames (Section 6.3).

6.1 Model Uncertainty

We have previously assumed that there is a unique conjecture about the behavioral model, while it may be more appropriate to assume that a regulator considers a number of different models possible. We therefore 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 might be uncertainty about the aspiration level of a satisficer, and one of the models in D could also be the rational agent.[14] 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. 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.

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 an observed data set is $DP(\Lambda) = \{(d, \succeq) \mid \Lambda \subseteq \bar{\Lambda}(d, \succeq)\}$. We again assume that $DP(\Lambda)$ is non-empty, i.e.,

[14] 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.

there is at least one conjecture that is not falsified by the data (see RS p. 377 for a discussion of multiple conjectures and their rejection). 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 to take into account that both model and welfare preference are unknown. In particular, 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 choice under frame f is at least as good as 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 the main analysis, we consider maximal data sets only and look for conditions under which $G(\bar{\Lambda}(d, \succeq))$ is non-empty for a true but unobservable pair $(d, \succeq)$. An immediate extension of Definition 2 could require identifiability of $\succeq$ in d. 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 very different normative implications (see e.g. Kőszegi and Rabin, 2008b; Bernheim, 2009), identifiability in the extended setting must aim at all aspects of the pair $(d, \succeq)$ that are normatively relevant.

Definition 5 *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 $d(\succeq, f) \neq d'(\succeq', f)$.*

Proposition 5 *$G(\bar{\Lambda}(d, \succeq))$ is non-empty if and only if $(d, \succeq)$ is virtually identifiable.*

The proof is similar to the proof of Proposition 1 and therefore omitted. 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 might not be able to predict the behavior of an agent with a different welfare preference $\succeq' \neq \succeq$. The property of virtual identifiability of $(d, \succeq)$ is clearly stronger than identifiability of $\succeq$ in d. For instance, we can have multiple models with identifiable preferences each, that, if considered jointly, do not have virtually identifiable model-preference pairs.[15] On the other hand, adding a rational agent to any

[15] As an example, let $m_X = 2$ so that $P = \{\succeq_1, \succeq_2\}$, and let $F = \{f_1, f_2\}$. Consider model d_1 given by $d_1(\succeq_1, f_1) = \succeq_1$, $d_1(\succeq_2, f_1) = \succeq_2$, $d_1(\succeq_1, f_2) = \succeq_2$, and $d_1(\succeq_2, f_2) = \succeq_1$, so that frame f_1 is non-distorting for both preferences while frame f_2 maps each preference into the other. Both preferences are identifiable. Now consider d_2 given by $d_2(\succeq_1, f_1) = \succeq_2$, $d_2(\succeq_2, f_1) = \succeq_1$, $d_2(\succeq_1, f_2) = \succeq_1$, and $d_2(\succeq_2, f_2) = \succeq_2$, so that the roles of f_1 and f_2 are reversed and both preferences are again identifiable. If $D = \{d_1, d_2\}$, no model-preference pair is virtually identifiable, because data sets can always be explained by the two different models with opposing welfare preferences.

given behavioral model with identifiable preferences preserves the property of virtually identifiable model-preference pairs.[16] Thus the possibility of agents being rational has no substantial impact on our previous analysis.

The analysis in Sections 4 and 5 can also be adapted to the case of model uncertainty. For instance, if each distortion function $d \in D$ satisfies the frame-cancellation property, then it follows immediately that no data set allows us to exclude any dominated frame. Applications include the uncertainty about a satisficer's (fixed) aspiration level. With virtually identifiable model-preference pairs, on the other hand, elicitation procedures now generate sequences of expanding data sets with the goal of learning about both preferences and models.

6.2 A Theory-Free Approach

We can go one step further and try to dispense with any conjecture about the behavioral model. Instead of following our theory-based approach to behavioral welfare economics, we could work with the purely choice-based approach by Bernheim and Rangel (2009). In fact, we can easily adapt our definition of the binary nudging relation and evaluate the frame-induced choices based on the weak unambiguous choice relation R' (Bernheim and Rangel, 2009, p. 60) rather than on a set of welfare preferences. Formally, a generalized choice situation (GCS) consists of a set of alternatives $S \subseteq X$ and a frame $f \in F$, and a choice correspondence describes the chosen alternatives for each GCS that we have observed. For better comparability, let us assume that the observed choice has always been a unique alternative $C(S, f) \in S$. To eliminate all traces of non-choice-based theories about mistakes, let us also assume that all the observed GCSs are welfare-relevant. Now consider two frames f and f' of which we know that they have a differential impact on behavior, i.e., we have observed two GCSs $(\bar{S}, f)$ and $(\bar{S}, f')$ with $C(\bar{S}, f) = x \neq y = C(\bar{S}, f')$. In line with our previous analysis, we could say that f is a weak unambiguous nudge over f' if $C(S, f)\, R'\, C(S, f')$ holds for all matching pairs (S, f) and (S, f') that we have observed. It follows immediately from the definition of R' that such a ranking is impossible. The mere fact that $C(\bar{S}, f) = x \neq y = C(\bar{S}, f')$ implies that neither $xR'y$ nor $yR'x$ holds, and hence neither of the two frames can be a weak unambiguous nudge over the other. If we worked with R^* (Bernheim and Rangel, 2009, p. 60) instead of R', we would obtain the statement that each frame is always a weak unambiguous nudge over every other frame. It follows that nudging is impossible without non-choice-based assumptions about decision mistakes, as already pointed out by Bernheim and Rangel (2009, p. 62).

[16] Let $D = \{d, d^R\}$, where d is the given model and d^R is the rational agent. Since each preference is identifiable in each model separately, we only need to check across models. Consider any $(d, \succeq)$ and $(d^R, \succeq')$ with $\succeq' \neq \succeq$. Let f be the non-distorting frame from $\succeq$ in d. Then $d(\succeq, f) = \succeq \neq \succeq' = d^R(\succeq', f)$.

6.3 Imperfectly Observable Frames

Finally, we have previously assumed that frames are perfectly observable and controllable by the regulator. Since a frame can be very complex in many contexts, this assumption is restrictive and deserves to be relaxed. For instance, consider a modified 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 description of 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. Perfect observation is captured by the finest partition $\Phi = \{\{f\} \mid f \in F\}$. 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 ϕ'. Thus a welfare preference $\succeq$ is consistent with Λ 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 Λ 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 ϕ 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 Φ 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 Φ_N. In continuation of our previous assumption, we suppose 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). 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 ϕ with ϕ' 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 Λ and ϕ. As before, let $[\phi]_\Lambda N(\Lambda)[\phi']_\Lambda$ if for each $(h, \succeq) \in H \times P(\Lambda)$ it holds that $c(d(\succeq, h(\phi)), S) \succeq c(d(\succeq, h(\phi')), S)$, for all non-empty $S \subseteq X$.

As in the main analysis we let $G(\Lambda) = \{\phi \mid [\phi]_\Lambda N(\Lambda)[\phi']_\Lambda, \forall \phi' \in \Phi_N\}$ and consider only maximal data sets to investigate the existence of an optimal nudge. An immediate extension of identifiability of $\succeq$ (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. This implies $\{\succeq, \succeq'\} \subseteq P(\bar{\Lambda}(\succeq))$ and makes it impossible to find an optimal nudge.[17] Our notion of identifiability in the setting with imperfectly observable frames must therefore ensure that the maximal data set reveals itself as maximal.

Definition 6 *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 $f' \in \phi$.*

Proposition 6 *$G(\bar{\Lambda}(\succeq))$ is non-empty if and only if $\succeq$ is potentially identifiable.*

The proof is again omitted. 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 $f' \in \phi$. We use the term potential identifiability here, because, while observation of $\bar{\Lambda}(\succeq)$ now allows us to conclude that this data set is maximal and $\succeq$ is the true welfare preference,

[17] As an example, let $m_X = 2$ so that $P = \{\succeq_1, \succeq_2\}$. Let $F = \{f_1, f_2, f_3\}$ and $\Phi = \{\phi_1, \phi_2\}$ with $\phi_1 = \{f_1, f_2\}$ and $\phi_2 = \{f_3\}$. Suppose $d(\succeq_1, f_1) = \succeq_1$, $d(\succeq_1, f_2) = \succeq_2$, $d(\succeq_1, f_3) = \succeq_1$, $d(\succeq_2, f_1) = \succeq_2$, $d(\succeq_2, f_2) = \succeq_2$, and $d(\succeq_2, f_3) = \succeq_1$. The maximal data sets are $\bar{\Lambda}(\succeq_1) = \{(\succeq_1, \phi_1), (\succeq_2, \phi_1), (\succeq_1, \phi_2)\}$ and $\bar{\Lambda}(\succeq_2) = \{(\succeq_2, \phi_1), (\succeq_1, \phi_2)\}$, such that $\bar{\Lambda}(\succeq_2) \subset \bar{\Lambda}(\succeq_1)$ and $P(\bar{\Lambda}(\succeq_2)) = \{\succeq_1, \succeq_2\}$.

there is still no guarantee that we will ever arrive at $\bar{\Lambda}(\succeq)$. An appropriately redefined elicitation procedure might impose a set of frames ϕ multiple times on the agent, but a specific element $f \in \phi$ still does not realize. This is in contrast to the case of observable frames, where a maximal data set can always be collected in exactly m_F steps.

7 Conclusions

Throughout the paper, we have taken the usual revealed-preference perspective for a single agent, which, aside from its methodological justification, 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, individual-specific learning and nudging is achievable, for instance by relying on cookies. However, our results also speak to the problem of learning and nudging for a population 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 distribution, would allow us to combine observations of different agents into a single data set, with the goal of learning about their welfare preferences (see Goldin and Reck, 2015). On the nudging stage, finding one optimal frame will be even more difficult for heterogeneous agents than for a single agent, and may require suitably adjusted criteria of optimality.

In fact, the analysis in this paper uses no prior information about the agent's welfare preference. If such information was available, the set of possible welfare preferences P might be smaller from the outset. More sophisticated information could be captured by a probability distribution on P. Our questions could then be asked in a probabilistic sense. What is the regulator's updated belief about the agent's welfare preference after having observed a behavioral data set? How likely is frame f a successful nudge over frame f'? It would also become possible to compare elicitation procedures by their expected running times. The next logical step would be to introduce cardinal utilities, which allows measuring and aggregating the severity of mistakes that frames induce. All these extensions come at the cost of substantially stronger assumptions, but make it possible to evaluate a nudge based on criteria such as the probability of optimality or the expected welfare. We leave these approaches to future research.

References

Bernheim, B. (2009). Behavioral welfare economics. *Journal of the European Economic Association*, 7:267–319.

Bernheim, B. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124:51–104.

Camerer, C. F., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". *University of Pennsylvania Law Review*, 151:1211–1254.

Caplin, A. and Martin, D. (2012). Framing effects and optimization. Mimeo.

De Clippel, G. and Rozen, K. (2014). Bounded rationality and limited datasets. Mimeo.

Goldin, J. (2015). Which way to nudge? uncovering preferences in the behavioral age. *Yale Law Journal*, forthcoming.

Goldin, J. and Reck, D. (2015). Preference identification under inconsistent choice. Mimeo.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, 38:635–645.

Kőszegi, B. and Rabin, M. (2007). Mistakes in choice-based welfare analysis. *American Economic Review, Papers and Proceedings*, 97:477–481.

Kőszegi, B. and Rabin, M. (2008a). Choice, situations, and happiness. *Journal of Public Economics*, 92:1821–1832.

Kőszegi, B. and Rabin, M. (2008b). Revealed mistakes and revealed preferences. In Caplin, A. and Schotter, A., editors, *The Foundations of Positive and Normative Economics*, pages 193–209. Oxford University Press, New York.

Kőszegi, B. and Szeidl, A. (2013). A model of focusing in economic choice. *Quarterly Journal of Economics*, 128:53–104.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. (2012). Revealed attention. *American Economic Review*, 102:2183–2205.

Rubinstein, A. and Salant, Y. (2006). A model of choice from lists. *Theoretical Economics*, 1:3–17.

Rubinstein, A. and Salant, Y. (2008). Some thoughts on the principle of revealed preference. In Caplin, A. and Schotter, A., editors, *Handbooks of Economic Methodologies*, pages 115–124. Oxford University Press, New York.

Rubinstein, A. and Salant, Y. (2012). Eliciting welfare preferences from behavioural data sets. *Review of Economic Studies*, 79:375–387.

Salant, Y. and Rubinstein, A. (2008). (A,f): Choice with frames. *Review of Economic Studies*, 75:1287–1296.

Spiegler, R. (2015). On the equilibrium effects of nudging. *Journal of Legal Studies*, forthcoming.

Sunstein, C. (2014). *Why Nudge? The Politics of Libertarian Paternalism.* New Haven: Yale University Press.

Thaler, R. and Sunstein, C. (2003). Libertarian paternalism. *American Economic Review, Papers and Proceedings*, 93:175–179.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* New Haven: Yale University Press.

A Proof of Proposition 3

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 $X\backslash f$ of the search engine.

Case 1: k even. We first construct an elicitation procedure e and then show that it is optimal. Let $e(\varnothing) = f_1$ 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 $X\backslash f_1$. 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 $X\backslash f_1$ 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 $X\backslash f_2$. This construction is applied to all the data sets Λ_1 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 Λ_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 Λ_1 has already revealed that each $b_1, \ldots, b_k$ is preferred by $\succeq$. Now suppose that $b_k \succeq_2 a_1$ holds. We then know that $b_k \succeq a_1$ and thus $T_k(\succeq) = \{b_1, \ldots, b_k\}$. Otherwise, if $a_1 \succeq_2 b_k$ 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 $b_{k-1} \succeq_2 a_2$ we know that $b_{k-1} \succeq a_2$ and thus $T_k(\succeq) = \{b_1, \ldots, b_{k-1}, a_1\}$; otherwise, if $a_2 \succeq_2 b_{k-1}$ 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 $a_{k/2} \succeq_2 b_{k/2+1}$, 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 $X \backslash f_2$.

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, except that the iteration might arrive at $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$, 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 $X \backslash f_2$.

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 $a_{(k-1)/2} \succeq_2 b_{(k+1)/2+1}$. 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 $X \backslash f_2$. 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 $e(\varnothing) = f_1 = \{a_1, \ldots, a_k\}$ and $X \backslash f_1 = \{b_1, \ldots, b_k\}$, 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 $b_{k-l+1} \in X \backslash f_2$ 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

$$\begin{aligned}
\succeq'\text{:}&\ \ a_1 \ldots a_{l-1}\, b_1 \ldots b_{k-l}\, b_{k-l+1}\, a_l\, a_{l+1} \ldots a_k\, b_{k-l+2} \ldots b_k,\\
\succeq''\text{:}&\ \ a_1 \ldots a_{l-1}\, b_1 \ldots b_{k-l}\, a_l\, b_{k-l+1}\, a_{l+1} \ldots a_k\, b_{k-l+2} \ldots b_k.
\end{aligned}$$

If $l = k$, let

$$\begin{aligned}
\succeq'\text{:}&\ \ a_1 \ldots a_{k-1}\, b_1\, a_k\, b_2 \ldots b_k,\\
\succeq''\text{:}&\ \ a_1 \ldots a_{k-1}\, a_k\, b_1\, b_2 \ldots b_k.
\end{aligned}$$

For the two constructed welfare preferences $\succeq'$ and $\succeq''$, the elicitation procedure first generates the above described data set Λ_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 $n > 2$.