Procedurally rational framing effects

Abstract

Framing effects are often taken as paradigmatic examples of human irrationality. The irrationality of framing effects is then used in debunking arguments against moral and philosophical intuitions. I argue that many framing effects are procedurally rational in the sense that they result from rational processes of practical inquiry. I make this argument through case studies of category-based choice, list-based choice, and salience-driven decisionmaking. I conclude by showing how the procedural rationality of framing effects can be used to resist framing-based debunking arguments against moral and philosophical intuitions.

1 Introduction

One of the most prominent findings in human judgment and decisionmaking is our vulnerability to framing effects (Bermúdez 2020; Levin et al. 1998; Tversky and Kahneman 1981). We are risk-averse when outcomes are presented as gains, but risk-seeking when outcomes are reframed as losses (Tversky and Kahneman 1981). We prefer objects whose attributes are positively framed, such as 90% lean beef, to the same objects when their attributes are negatively framed, such as 10% fat beef (Levin and Gaeth 1988). Framing effects are pronounced enough in human activity that they have begun to emerge in large language models (Binz and Schulz 2023; Jones and Steinhardt 2022).

Framing effects have been alleged to problematize many philosophical projects. Most prominently, they are taken to show that moral intuitions (Horowitz 1998; Sinnott-Armstrong 2008) and other philosophical intuitions (Machery 2017; Weinberg et al. 2010) are unreliable, because those intuitions are subject to framing effects.[1] The existence of framing effects has been taken to challenge notions such as informed consent (Chwang 2016; Hanna 2011) and moral expertise (Horvath and Wiegmann 2020; Schwitzgebel and Cushman 2012).[2] More broadly, framing effects are used to justify interventions such as

[1] For pushback see Demaree-Cotton (2016); Dreisbach and Guevara (2019) and to some extent May (2018).

[2] For pushback see Cohen (2013) and Director (2024) on informed consent and Horvath (2010) and Williamson (2011) on moral expertise.

nudging (Sunstein 2014; Thaler and Sunstein 2008) and debiasing (Fischoff 1982; Larrick 2004) aimed at correcting or working around judgments vulnerable to framing effects.

In response to these projects, many philosophers have aimed to rationalize framing effects. It has been argued that some frames are normatively inequivalent and hence merit different responses (Dreisbach and Guevara 2019; Kamm 1998), that framing is appropriate in non-intensional contexts whose complexity cannot be adequately captured by a single frame (Bermúdez 2020, 2022), or that framing arises from rational learning and inference (Horne and Livengood 2015; Gigerenzer 2018). Many of these strategies face challenges (Horowitz 1998; Sinnott-Armstrong 2008; Stein 1996) and none are universally applicable, creating a need for supplementary rationalizing strategies.

This paper aims to show how many framing effects can be procedurally rational, in the sense of resulting from rational processes of practical inquiry. Drawing on the frame-based choice model of Yuval Salant and Ariel Rubinstein (2008), I characterize the notion of procedurally rational framing effects (Section 2.2). I then give three examples of how framing effects can be procedurally rational because they emerge from core features of human cognition that are rationally indispensable to creatures like us in many circumstances.

Section 3 shows how framing effects arise from category-based choice, in which agents use what they have learned about categories of options to guide choice and updating. Section 4 shows how framing effects arise when agents with limited memory and computational power examine options in a particular order, rather than all at once. Section 5 shows how agents who use salience to draw attention to potentially important features of options can be vulnerable to framing effects due to salience manipulations.

Section 6 draws out the immediate lesson of this argument: framing effects can arise from rational processes in which the very same features that drive framing effects are not irrational artifacts but rather rationally essential features of human cognition. Section 7 uses this finding to develop two challenges for recent attempts to use framing effects to debunk moral intuitions.

2 Procedurally rational frame-based choice

This section characterizes choice with frames (Section 2.1) and procedurally-rational frame-based choice (Section 2.2).

2.1 Frame-based choice

Fix a finite set X of alternatives. Classically, a choice problem is a nonempty subset A of alternatives from X. Agents select alternatives according to a choice function $c : \mathcal{P}(X) \setminus \emptyset \longrightarrow X$ returning a chosen alternative from each choice problem, so that $c(A) \in A$ for all nonempty $A \subseteq X$.

Following Yuval Salant and Ariel Rubinstein (2008), we can model choice in framed decision problems by specifying a finite set $\mathcal{F}$ of frames. In this setting, a choice problem is a pair (A, f) where A is a nonempty subset of X and $f \in \mathcal{F}$ is a frame. A choice function $c : (\mathcal{P}(X) \setminus \emptyset) \times \mathcal{F} \longrightarrow X$ chooses alternatives in a frame-dependent way, so that $c(A, f) \in A$. Framing effects are defined as follows:

> Choice function c **exhibits framing effects** if there exists a nonempty set $A \subseteq X$ and frames $f, f' \in \mathcal{F}$ with $c(A, f) \neq c(A, f')$.

A classic example of framing effects arises in the Asian Disease Problem (Tversky and Kahneman 1981).

Consider two interventions aimed at preventing an Asian disease.

> ($\mathbf{x_1}$, **gain**) 200 people will be saved.
>
> ($\mathbf{x_2}$, **gain**) There is a 1/3 chance that 600 people will be saved, and a 2/3 chance that 0 people will be saved.

Tversky and Kahneman (1981) found that most participants expressed a preference for (x_1, gain) over (x_2, gain).

Now suppose that the same lotteries are expressed in terms of lives lost:

($\mathbf{x_1}$, **loss**) 400 people will die.

($\mathbf{x_2}$, **loss**) There is a 1/3 chance that 0 people will die, and a 2/3 chance that 600 people will die.

Tversky and Kahneman (1981) found that most participants expressed a preference for (x_2, loss) over (x_1, loss). This is a framing effect with $A = \{x_1, x_2\}$ understood as lotteries over resulting lives and $\mathcal{F} = \{\text{gain, loss}\}$.

While some have defended the rationality of participants' behavior (Dreisbach and Guevara 2019; Gigerenzer 2018), framing effects in the Asian disease problem are taken by many as a paradigmatic example of irrationality, including by some defenders of rational framing effects (Bermúdez 2020). However, even if the Asian disease problem is taken to be an example of an irrational framing effect, it is important to avoid overgeneralizing on salient examples of irrational framing effects. My claim in this paper is that many framing effects are potentially unlike framing effects in the Asian disease problem in that they are procedurally rational. The next task is to say what this means.

2.2 Procedurally rational frame-based choice

Humans are bounded agents. We are bounded internally our cognitive architecture and externally by our environment. The study of bounded rationality asks what rationality requires of bounded agents (Gigerenzer and Selten 2001; Thorstad 2024; Viale 2021).

Herbert Simon (1976) held that the fundamental turn in the study of bounded rationality is the turn from substantive to procedural rationality. Because many paradigmatic bounds are felt strongly as bounds on the processes we can execute, theories of bounded rationality take a procedural lens which emphasizes rational constraints on processes.

This paper follows an interpretation of the procedural turn due to David Thorstad (forthcoming). On this interpretation, theories of *substantive rationality* ask questions about the rationality of attitudes such as preference, intention, belief and credence. Theories of *procedural rationality* move a level up, asking questions about the rationality of processes of inquiry which produce and modify attitudes.

A strong reading of procedural rationality joins procedural rationality with indirect normative theory to link the rationality of attitudes to the rationality of processes which produce them:

> **(Strong Procedural Rationality)** For all agents S, times t and attitudes A, S's attitude A is rational at time t if and only if A was produced by a rational process of inquiry at or before time t.

For example, we might hold that a process of theoretical inquiry is rational just in case it is sufficiently reliable and that a doxastic attitude is rational just in case it results from a sufficiently reliable process. On this view, the paper's claim is:

> **(Strong Procedural Rationality of Framing Effects)** For some agent S, time t, deliberative process P, and attitude A:
>
> > **(Formation)** S uses P to form A at t.
> >
> > **(Rational process)** It is rationally permissible for S to use P at t.
> >
> > **(Rational attitude)** S's attitude A is rational at t.
> >
> > **(Framing)** P exhibits framing effects.

For example, we might claim that reliable processes of inquiry sometimes exhibit framing effects, but that attitudes formed by reliable processes are rational nonetheless.

My own sympathies lie with direct normative theorists, who apply a strict level separation (Thorstad 2024) between the rationality of attitudes and the rationality of processes which produce them. This allows the rationality of processes to come apart from the rationality of attitudes. To say that framing effects can result from rational deliberative processes is then not yet to say anything about the rationality of the resulting attitudes. On this reading, the paper defends a claim about the rationality of processes that exhibit framing effects:

(Procedural Rationality of Framing Effects) For some agent S, time t, deliberative process P, and attitude A:

> **(Formation)** S uses P to form A at t.
>
> **(Rational process)** It is rationally permissible for S to use P at t.
>
> **(Framing)** P exhibits framing effects.

Strong Procedural Rationality of Framing Effects entails Procedural Rationality of Framing Effects. As such, readers sympathetic to indirect normative theories are welcome to adapt the arguments in this paper to support Strong Procedural Rationality of Framing Effects. However, I argue only for Procedural Rationality of Framing Effects.

My strategy is to exhibit three plausibly rational deliberative processes, to argue that these processes are rational when used in appropriate circumstances, and to show how these processes exhibit framing effects. For generality, the processes are not randomly selected, but rather respond to essential features of human cognition such as categorization (Section 3), complexity limitations (Section 4) and salience-driven processing (Section 5). In each case, we will see that processes which exhibit framing effects respond to core features of human cognition that are broadly beneficial, and whose benefits could not be clearly replicated without vulnerability to framing effects. This will help us to see how a wide range of framing effects arise from rational processes, so that Procedural Rationality of Framing Effects is not an isolated phenomenon but rather a broad claim about many important framing effects.

3 Categorization

One of the central features of human cognition is conceptually-driven categorization. Categorization facilitates tasks such as learning, inference, understanding, explanation, planning, and communication (Medin and Heit 1999). As such, it is often helpful to cognize directly in terms of categories instead of, or prior to considering the objects that fall within them. For example, we decide to dine at an Italian restaurant before selecting a particular restaurant, or that tonight's entertainment will be a show before selecting a particular show.

This section considers a leading model of category-based choice (Section 3.1) due to Paola Manzini and Marco Mariotti (2012). We will see that category-based choosers are vulnerable to framing effects, but that this vulnerability may not present a decisive objection to the rationality of category-based choice (Section 3.2).

3.1 Category-based choice

In *category-based choice* (Manzini and Mariotti 2012), agents divide objects into categories. They rule out some categories of objects, then choose among remaining objects without considering categories (Figure 1).

Suppose you are shopping for cars. There are a finite set X of cars that you could buy. You group X into a set of categories $\mathcal{C} \subseteq \mathcal{P}(X)$. For simplicity, this model ignores conceptual structure and identifies categories extensionally with the objects they contain. Perhaps $\mathcal{C}$ contains categories such as SEDAN, HATCHBACK, PICKUP, VAN and SPORTS CAR.

Let us suppose that if you were to exhaustively compare and evaluate each car, you would develop complete and asymmetric preferences $\succ_2$ on all available cars X. But that is not always feasible and makes poor use of existing knowledge about categories. Before evaluating cars, you eliminate some categories of cars that are unlikely to be chosen.

You do this by developing an asymmetric partial ordering $\succ_1$ on categories $\mathcal{C}$, where $C_1 \succ_1 C_2$ indicates that you find category C_1 better than category C_2. For example, perhaps you do not care for pickups, vans and sports cars, so that SEDAN $\succ_1$ PICKUP, VAN, SPORTS CAR and HATCHBACK $\succ_1$ PICKUP, VAN, SPORTS CAR. However, you're not ready to pronounce between sedans and hatchbacks. At the first stage, you eliminate all elements of dispreferred categories, leaving a winnowed set $X' = \{x : (\nexists C_1, C_2 \in \mathcal{C})(C_1 \succ C_2 \wedge x \in C_2)\}$. In this case, X' = SEDAN $\cup$ HATCHBACK is the set of sedans and hatchbacks.

Next, you fully evaluate all remaining options and choose between them according to

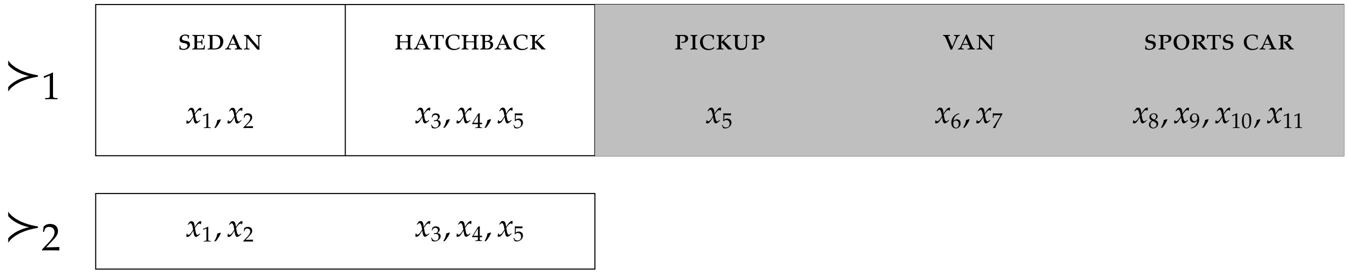

Figure 1: Category-based choice: Agents select categories according to $\succ_1$ then choose among elements of remaining categories to maximize $\succ_2$.

$\succ_2$. In this case, you choose the best sedan or hatchback on the market. More generally, you choose x^* to be the unique $\succ_2$-maximal element of X', in the sense that for all $x' \in X' \setminus \{x^*\}, x^* \succ_2 x'$.

Category-based choice is a promising way to choose among large menus of differentiated options when categories carry relevant information. Indeed, category-based choice is close to the method that I recently used to settle on buying a Toyota Corolla.

We can regard category-based choice as an instance of frame-based choice where a frame f is a pair $(\mathcal{C}, \succ_{\mathcal{C},1})$ of a categorization $\mathcal{C}$ and an asymmetric betterness relation $\succ_{\mathcal{C},1}$ on categories. Framing effects occur when for some choice set X, categorizations $\mathcal{C}, \mathcal{C}'$ and betterness relations $\succ_{\mathcal{C},1}, \succ_{\mathcal{C}',1}$ choice differs between $(X, \mathcal{C}, \succ_{\mathcal{C},1})$ and $(X, \mathcal{C}', \succ_{\mathcal{C}',1})$. This happens when the $\succ_2$-maximal elements of $X'_{\succ_{\mathcal{C},1}}$ and $X'_{\succ_{\mathcal{C}',1}}$ differ.

3.2 Framing effects in category-based choice

Suppose that Toyota has recently released a van x_V which is so high-quality that I would prefer it to a Corolla if I were to deeply consider their comparative merits. In our example above, x_V is not chosen because it is eliminated in the first round of choice.

However, I might have used an alternative categorization where categories $\mathcal{C}'$ are car brands such as Toyota, Ford, BMW and Lexus. Perhaps I prefer Toyota and Ford to the other brands, so that ruling out dominated brands according to $\succ_{\mathcal{C}',1}$ leaves the options $X' =$ Toyota $\cup$ Ford. Van x_V is $\succ_2$-maximal in X', and hence is chosen. This is a framing effect because choice differs between $(X, \mathcal{C}, \succ_{\mathcal{C},1})$ and $(X, \mathcal{C}', \succ_{\mathcal{C}',1})$.

If we wanted to argue that vulnerability to framing effects makes this and similar processes of category-based choice irrational, we would have to make one of two claims. First, we could claim that:

> **(Irrationality of Category-Based Choice)** Category-based choice is always irrational.

Alternatively, we could claim that:

> **(Frame-Independence of Rational Category-Based Choice)** Rational processes of category-based choice should be immune from framing effects.

Both claims face challenges.

Irrationality of Category-Based Choice confronts the Problem of Forgone Advantages of Categorization:

> **(Problem of Forgone Advantages of Categorization)** Forgoing category-based choice deprives agents of the cognitive advantages of categorization.

We saw above that categorization confers important cognitive advantages. Categorization enables efficient winnowing of options so that scarce computational and attentional resources can be allocated where they are most likely to be useful. Categorization enables learning and generalization at the level of categories, which leads to improved inference and decisionmaking. Categorization also facilitates communication and justification.

These are not trivial advantages. They are the reasons why categorization plays a central role in human cognition. Blanket avoidance of category-based choice deprives agents of the advantages of categorization. Agents then risk wasting scarce attentional and computational resources, failing to acquire or bring to bear knowledge about categories, and making decisions that are difficult to recall and communicate. In some situations, there may be alternative means to secure the benefits of categorization. For this reason, I do not claim that category-based choice is always rational. But neither should we jump to the opposite conclusion and assume that the benefits of category-based choice can always be replicated or outweighed.

By contrast, Frame-Independence of Rational Category-Based Choice urges us to slice categories so finely that they will be in principle invulnerable to framing effects. I am not sure if this could be done, but if it could it would face the Problem of Forgone Advantages of Generality:[3]

> **(Problem of Forgone Advantages of Generality)** Finely individuating categories deprives agents of the cognitive advantages of general categories.

Thin categories have few instances, making them hard to learn about, generalize upon, or communicate with, depriving them of many of the advantages claimed for categorization. Thin categories are also provably suboptimal in many settings.

Leading models of optimal categorization suggest that the optimal number of categories into which objects should be partitioned grows slowly – typically, more slowly than the number of objects partitioned (Anderson 1990; Mohlin 2014; Sandborn et al. 2006). One way to understand this result is that agents face a bias-variance tradeoff in categorization (Geman et al. 1992; Mohlin 2014). Categories are predictive tools allowing us to anticipate the properties of novel objects by considering the properties of other category members. Predictive error is driven both by the *bias* of a prediction rule, understood as its tendency to systematically favor certain directional predictions, and by its *variance*, a measure of overfitting. Increasing the number of categories reduces bias by reducing the directional constraints imposed by categorization schemes. However, increasing the number of categories increases variance because the boundaries of smaller categories are more likely to be overfitted to artifacts of observed data. As a result, using a relatively small number of more general categories often helps agents to make more accurate predictions than they would by using a large number of thinly-sliced categories.

[3] Certainly this is impossible without restriction on the space of admissible categorizations and betterness relations on categories, unless we are to restrict ourselves to one-element categories, which are hardly categories at all.

If this is right, then there is no general route to the conclusion that framing effects from category-based choice must always be the result of irrational processes. Pressing the Irrationality of Category-Based Choice confronts the Problem of Forgone Advantages of Categorization. Categorization is useful, and we should not always forgo its advantages. Opting instead for Frame-Independence of Rational Category-Based Choice faces the Problem of Forgone Advantages of Generality. General categories are useful, and we should not always forgo their advantages. This suggests that category-based choice is a first example of procedurally rational framing effects.

4 List-based choice

Humans do not evaluate large groups of options all at once. Rather, we consider options in a specific order. This order may be exogenously given, as when a real estate agent chooses the order in which to present her client with homes, or endogenously given, when agents choose orderings for themselves. Agents who decide in this way choose not from unordered sets of options, but rather from ordered lists (Rubinstein and Salant 2006; Salant 2011).

This section argues that rational processes of list-based choice sometimes exhibit framing effects. After introducing list-based choice and characterizing framing effects in this context (Section 4.1) we will consider list-based choice within a specific cognitive architecture: finite automata (Section 4.2). We will see that framing effects are unavoidable for complexity-limited agents within this architecture (Section 4.3). By examining how framing effects arise from optimal responses to complexity constraints, we will see how a process' vulnerability to framing effects can be construed not as an irrational artifact, but rather as a means of learning during list-based choice (Section 4.4).

4.1 Lists and frames

Given a finite set A, a *list* on A is a bijection $L : A \longrightarrow \{1, 2, \ldots, |A|\}$. Lists determine the order in which options will be examined for choice. We can understand list-based choice as an instance of frame-based choice by taking frames to be lists and decision problems to be pairs (A, L) of alternative sets A and lists L on A. Framing effects occur when for some set A of alternatives and lists L, L' on A, the agent chooses differently in (A, L) and (A, L').

In this case, framing effects are *ordering effects* (Horne and Livengood 2015; Li and Epley 2009; Mantonakis et al. 2009). Items are evaluated differently based on the order in which they are presented. For example, agents may exhibit a *primacy effect* in which early list items are more likely to be chosen (Guney 2014; Miller and Krosnick 1998). Or they may exhibit a *recency effect* in which later list items are more likely to be chosen (Bjork and Whitten 1974; Guney 2014).

It may seem obvious that rational decision procedures will eliminate ordering effects. Indeed, ordering effects are among the most common framing effects used to cast doubt on moral intuitions (Horne and Livengood 2015; Swain et al. 2008; Wiegmann et al. 2012). But it is surprisingly difficult to eliminate ordering effects.

4.2 Finite automata

A good way to study the persistence of ordering effects is to study the complexity of choice procedures that avoid them. To do this, it is often helpful to fix a specific cognitive architecture in which the complexity of choice procedures can be precisely compared.

A popular choice is to consider *finite automata* (Banovetz and Oprea 2023; Rabin and Scott 1959; Rubinstein 1986). A finite automaton takes lists L from some domain X as input. It returns a list item l from L as output. Finite automata are specified as follows.[4]

An automaton has a finite set S of possible states. One state s_0 is designated as the initial state. Automata consider a list L by reading the list elements one at a time. After

[4] This formulation follows Salant (2011) for continuity with later formal results.

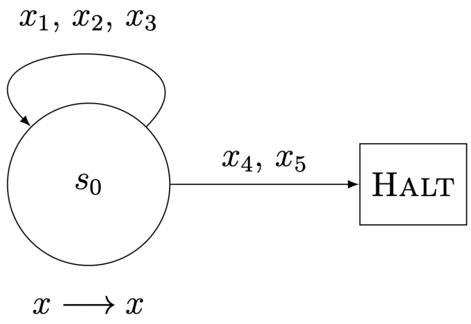

Figure 2: Satisficing

considering some element l, the automaton may shift its internal state or halt. This is done according to a *transition function* $\tau : S \times X \longrightarrow S \cup \{\textsc{halt}\}$ which considers the current state s and the current list item l being considered. Agents make a decision when they have halted or reached the end of the list. At this time, choice depends on the current state and the last examined list item according to a *selection function* $\sigma : S \times X \longrightarrow X$.

Consider a simple one-state automaton given by Figure 2. The automaton takes as inputs lists from a five-element domain $X = \{x_1, x_2, x_3, x_4, x_5\}$. It has a single state s_0. The transition function τ is represented by arrows between states. In this case, the agent remains in her current state when examining x_1, x_2 or x_3 but HALTs when examining x_4 or x_5. On halting, she chooses the last item examined, or the last list item if she reaches the end of the list without halting. This choice process is represented diagrammatically below her state.

We can study the complexity of choice rules by studying the complexity of automata which implement them. In this case, a choice rule is a function $c : \mathcal{L}(X) \longrightarrow X$ taking as input lists L on subsets of X and returning chosen items $c(L)$ from L. Finite automaton a on X *implements* a choice function c on X if for all lists $L \in \mathcal{L}(X)$, a outputs $c(L)$ on input L.

For example, let $X = \{x_1, x_2, x_3, x_4, x_5\}$. The automaton in Figure 2 implements *satisficing* with x_4, x_5 considered to be satisfactory. It examines list elements until encountering a first satisfactory element x_4 or x_5, then chooses this element.

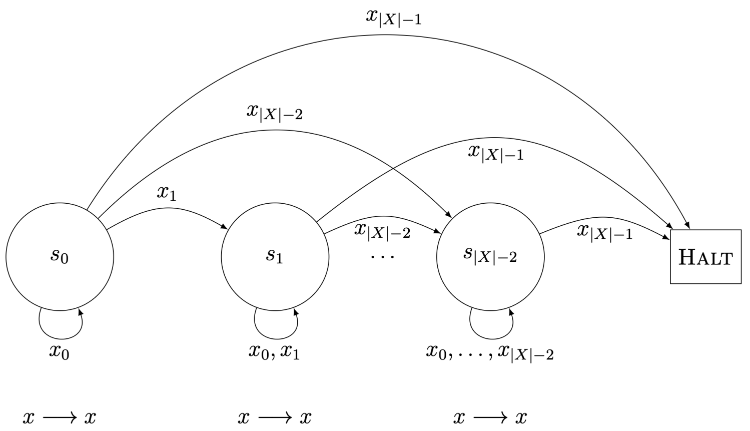

Figure 3: Utility maximization

4.3 Complexity and order invariance

The most popular complexity criterion for finite automata is *state complexity*: the number of states needed to implement them. Figure 2 shows that satisficing has minimal state complexity 1. And in fact satisficing is the unique choice function with minimal state complexity.[5]

By contrast, *utility maximization* ranks options $x \in X$ according to some real-valued utility function u and returns the utility-maximal element of any list.[6] One finite automaton implementing utility maximization is given by Figure 3. This automaton has state complexity $|X| - 1$, one fewer than the number of available options. With few options on the market, this may not be a problem. But if X is a large set, such as the cars on offer by local dealers, $|X| - 1$ may be a very large number of states indeed. We may hope that utility maximization could be implemented with fewer than $|X| - 1$ states. But this is not so. Utility maximization has state complexity $|X| - 1$.[7]

[5] To see this, let $u(x) = 1$ for any option chosen in s_0 and $u(x) = 0$ for the remaining options, then pick a satisficing threshold of $t = 1/2$.

[6] Throughout this section I assume that no two options have the same utility. Removing this assumption would reduce the state complexity of utility maximization by the number of tied elements and allow ordering effects among elements tied for best, though many choice rules permit randomization between best elements.

[7] This is immediate from the fact that rules of complexity $< |X| - 1$ exhibit order effects, whereas utility maximization does not.

Satisficers exhibit order effects. Our satisficer chooses x_4 from the list $x_1x_4x_5$ but chooses x_5 from the list $x_1x_5x_4$. By contrast, utility maximizers do not exhibit order effects. Faced with a list L, they always choose the u-maximizing element of L, no matter the order in which L is presented.

Say that a choice function c from lists is *order invariant* if for any list L and list permutation σ, $c(L) = c(\sigma(L))$. Choice functions exhibit ordering effects when they fail to be order invariant. It would be nice if order invariance could be purchased more cheaply than utility maximization. But in fact this is not the case. We can prove that for any set X, all order-invariant choice rules have state complexity of at least $|X| - 1$, the complexity of utility maximization.[8]

In this sense, when maximization is infeasible, then so is order invariance. Only agents who can spare the expense of full-on utility maximization can avoid order effects. This suggests two conditions under which agents might rationally use processes which exhibit order effects.

First, some agents are unable to implement processes of sufficient complexity to avoid order effects. Because ought implies can, such agents cannot be required to implement order-invariant processes.

Second, agents have limited cognitive resources. As a result, even if agents are able to implement complex order-invariant processes, this may come at the cost of cognitive resources that are needed for other judgments and decisions. Opting for order-invariance in one decision may not be worth the cost to other decisions. Further, although resource-limited agents can often implement order-invariant processes in some problems, they cannot do so in all decision problems. This suggests that it is impossible for many agents to implement order-invariant processes throughout their cognitive lives.

[8] See Salant (2011), Theorem 1.

4.4 Optimal processes and order effects

We can gain further insight into the causes of order effects by seeing how they emerge from optimal choice by complexity-limited agents. In this setting, we will see that order effects emerge not as irrational artifacts, but rather as a means of improving decisions through learning during list-based choice.

Suppose that an agent can spare more than enough complexity to implement satisficing, but not enough to implement utility maximization. That is, she must pick a choice rule with state complexity k for $1 < k < |X| - 1$. What process should she use?

Let us suppose that lists are drawn randomly in the sense of being generated by a stationary process. That is, for some probability function P on X with full support, the first list element is drawn according to P. List generation halts with some fixed probability $c > 0$ and otherwise a second element is drawn according to P, continuing in this way until generation halts. This process induces a corresponding probability distribution over lists that the agent may be presented with.

The agent wants to pick a choice rule c maximizing expected utility given her uncertainty about lists $\mathbf{L}$ she might be presented with. That is, she wants to pick c to maximize

$$E[u(c(\mathbf{L}))] = \sum_{L} Pr(\mathbf{L} = L)u(c(L)) \tag{1}$$

such that c has state complexity $\leq k$. In this setting, we can show that the utility-maximizing process is k-phase satisficing (Salant 2011).

Informally, k-phase satisficers string together k different satisficing agents with increasingly demanding thresholds $t_0, t_1, \ldots, t_{k-1}$. They begin by implementing the least demanding threshold t_0 but can raise their standards to some t_i when they encounter high-value list elements, encouraging them to be more picky.

More formally (Figure 4), k-phase satisficers fix a sequence $a_0, \ldots, a_{k-1}$ of *pivotal alternatives* chosen to generate a sequence of strictly increasing thresholds $t_i = u(a_i)$. They have states $s_0, \ldots, s_{k-1}$ corresponding to the utility thresholds. When encountering a non-pivotal

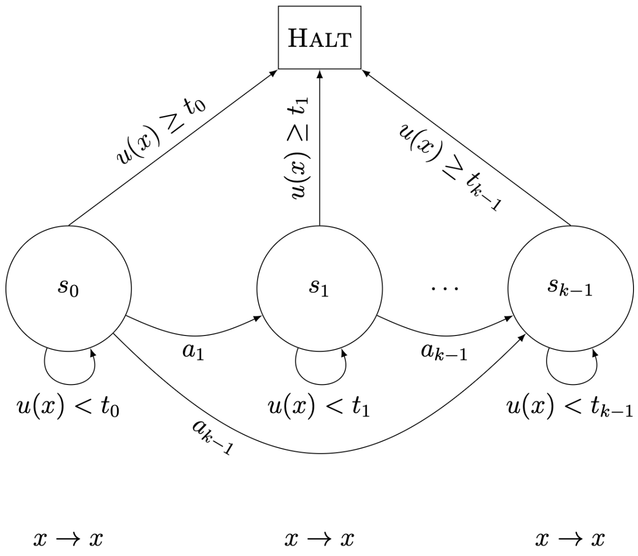

Figure 4: K-phase satisficing

alternative in state s_i, agents satisfice with threshold t_i, choosing the observed element x if $u(x) > u(t_i)$ and otherwise remaining in s_i. Pivotal alternatives are taken as invitations to raise satisficing thresholds, so that an agent encountering pivotal alternative a_j in state a_i shifts to s_j if $j > i$. If the agent reaches the end of the list, viewing alternative x in state s_i, she makes a forced choice between x and a_i to maximize u.

K-phase satisficers make better choices than satisficers because they are able to adjust their standards as they encounter each list. However, the very ability to adjust their standards generates ordering effects.

Suppose that $X = \{x_1, x_2, \ldots, x_{100}\}$ has a hundred elements with $u(x_i) = i$. Let us suppose that a k-phase satisficer with $k = 9$ treats $x_{10}, x_{20}, \ldots, x_{90}$ as pivotal alternatives. Now consider the list $x_{40}x_{15}x_{51}$. Here the agent sees x_{40}, taking this as an invitation to raise her utility threshold to $u(x_{40}) = 40$. She therefore rejects x_{15} and holds out for x_{51} with $u(x_{51}) = 51$.

Our agent is a good learner, but not a perfect one. From the list $x_{15}x_{40}x_{51}$ she chooses x_{15} because she has not had the opportunity to observe x_{40} and learn to raise her standards.

More generally, she shows *primacy effects*: moving elements forward in a list increases their likelihood of being chosen.

Once we understand why *k*-phase satisficers exhibit primacy effects and other ordering effects, much of their seeming irrationality disappears. *K*-phase satisficers exhibit ordering effects because observing high-value list elements helps them learn to be more selective. They cannot learn from what they have not seen, so that moving low-value elements to the front of a list or moving learning opportunities to the back of a list can make low-value elements more likely to be chosen. But this is not because *k*-phase satisficers intrinsically value serial position in lists. It is because *k*-phase satisficers are making optimal use of their complexity budget to learn to select high-value options.

Complexity-limited learning creates order effects, but it is nonetheless better than no learning in a very precise sense: it is expected utility maximizing. Although complexity-limited agents cannot entirely avoid order effects, they could use simpler processes which face fewer order effects. They could even, like simple satisficers, refuse to learn at all. In doing so, they would face fewer framing effects, but they would also make worse decisions.

4.5 Taking stock

This section introduced list-based choice and characterized framing effects as ordering effects in list-based choice (Section 4.1). Working with finite automaton models (Section 4.2) we saw that framing effects are unavoidable unless agents can spare enough complexity to implement full-on utility maximization (Section 4.3). When this is not possible, the optimal *k*-phase satisficing process shows framing effects for a good reason: ordering effects help *k*-phase satisficers to learn to select high-value options (Section 4.4). Both the general difficulty of avoiding framing effects and the learning benefits of optimal vulnerability to framing effects speak in favor of the rational permissibility of processes for list-based choice that exhibit framing effects.

5 Salience

It is increasingly held that we are transitioning from an information economy into an attention economy (Browne and Watzl 2025; Castro and Pham 2020; Simon 1971). In an information economy, decisionmakers are limited primarily by the availability of information. By contrast, in an attention economy we face a glut of potentially relevant information and are increasingly limited by our ability to attend to it.

Sometimes attention is regulated in a top-down, goal-directed manner. This type of attentional regulation is captured by theories such as rational inattention (Sims 2003, 2006) and sparsemax decisionmaking (Gabaix 2014). Other times, attention is directly responsive to features of stimuli in a bottom-up manner. This is the operation of salience, and is captured by distinct theories of salience-driven decisionmaking (Bordalo et al. 2022).

This section introduces salience and its effects on choice (Section 5.1), then considers a case study of attribute prominence in salience-driven decisionmaking (Section 5.2). We will see that the resulting framing effects have a good claim to be understood as the products of rational salience-driven decisionmaking (Section 5.3).

5.1 Salience and choice

There are many accounts of salience and its role in decisionmaking. For concreteness, this section follows a leading account due to Eric Knudsen (2007).

On this account, information affects decisionmaking by entering working memory, where it can be the target of deliberative processes that shape choice. Because there is not room for all information to enter working memory, neural representations of relevant information compete for space in working memory on the basis of their signal strength. Signal strength is affected by information quality and attentional signals, which can be top-down goal-driven signals or bottom-up salience signals.

Stimuli become salient due to features such as surprisingness, contrast with other stimuli, and prominence in perception (Bordalo et al. 2022). Information that is surprising, contrasting or prominent is more likely to be decision-relevant, urgent, or useful for avoiding rare but costly mistakes. As such, neural representations of salient stimuli receive increased signal strength, making them more likely to enter working memory.

This does not mean that salient stimuli bypass cognitive mechanisms of deliberation. Once information enters working memory, it must still be evaluated for quality and relevance, so that salient stimuli may be disregarded if they are irrelevant to decisionmaking. However, salient stimuli have the advantage that they are more likely to enter working memory where they can, if relevant, influence decisionmaking.

For concreteness, it will help to focus on a particular type of salient stimuli: stimuli that are prominent in perception. Stimuli may be prominent because they are visually central, as with the papers near the center of my desk, or when they are prominently emphasized, as with papers marked with sticky notes. Section 5.2 looks at the effects of prominence on a leading model of salience-driven decisionmaking.

5.2 An example: Prominence

The leading model of salience-driven decisionmaking is the model of Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer (2012; 2016; 2020). Let us consider a simple version of the model in which salience is driven only by prominence.

In this model, alternatives x have attributes $x_1, \ldots, x_n$. The true value of alternatives is additive over attributes, so that:

$$V(x) = \Sigma_i u(x_i). \tag{2}$$

However, some attributes are prominent and others are not. Let $\mathcal{P} \subseteq \{1, 2, \ldots, n\}$ contain the indices of prominent attributes.

Non-prominent attributes x_i are given reduced weight $r_i \in (0, 1)$ in decisionmaking, reflecting their reduced prominence.[9] This weight reflects the fact that information about

[9] Bordalo, Gennaioli and Shleifer endogenize r_i in terms of similarity-driven memory retrieval, but the details of r_i need not concern us here.

non-prominent attributes is less likely to enter working memory, where it can be weighed in decisionmaking. Agents then value options at

$$\hat{V}(x) = \sum_{i \in \mathcal{P}} u(x_i) + \sum_{j \notin \mathcal{P}} r_j u(x_j). \tag{3}$$

In the simplest case, agents face a choice between two alternatives $X = \{x, x'\}$ with the salient attributes $\mathcal{P}$ and weights r_i held constant across alternatives.

In this setting, a frame is a set $\mathcal{P} \subseteq \{1, 2, \ldots, n\}$ of prominent attribute indices and a set $\mathcal{R} = \{r_i\}_{i \notin \mathcal{P}}$ of weights. Framing effects arise when for some frames $(\mathcal{P}, \mathcal{R})$ and $(\mathcal{P}', \mathcal{R}')$ we have $c(X, \mathcal{P}, \mathcal{R}) \neq c(X, \mathcal{P}', \mathcal{R}')$. Such effects are not hard to generate.[10]

For example, in a discount store prices are prominently displayed. By contrast, in a luxury boutique signals of quality rather than prices are prominently displayed. This is often beneficial for consumers. Discount store shoppers gain improved access to information about prices, which is highly decision-relevant for them. Luxury boutique shoppers gain improved access to information about quality, which is highly decision-relevant for them.

However, manipulating prominence in either context may change what I buy. Let x be a low-price medium-quality option and x' be a high-price, high-quality alternative. In the discount store I choose x over x' because price is prominent. However, if the discount store owners were instead to make quality prominent, I might choose x' over x. This is a framing effect with choice set $\{x, x'\}$, standard prominence $\mathcal{P}$ placed on price, manipulated prominence $\mathcal{P}'$ placed on quality, and weights $\mathcal{R}$ extracted from my shopping habits. What should we make of the fact that salience-driven decisionmaking exposes me to framing effects of this kind?

[10] For other examples of salience-driven framing effects see Gerken (2022) and Whiteley (2022).

5.3 Evaluating salience-driven framing effects

Salience plays an important role in informing decisionmaking. Salience helps agents to quickly select relevant information and eliminate distracting stimuli (Knudsen 2007), guide inference and choice (Munton 2023) and shape priority structures among mental representations (Watzl 2017). Without salience, we would be slowed or incapacitated by irrelevant information, unable to make timely inferences and choices, and limited in our ability to prioritize among stored representations. For these reasons, salience plays a central role in human cognition.

It is true that features such as prominence, surprise, and contrast can be manipulated.[11] Salience manipulations affect the information that enters working memory and thereby frame decisionmaking. Salience manipulations are often beneficial, as in the tendency of discount retailers to make prices prominent. Some salience manipulations are not beneficial, such as the tendency of retailers to prominently display candy bars at the checkout line. But given the cognitive benefits of salience, neither the bare vulnerability to framing effects nor even the occasional vulnerability to malicious framing leads directly to an argument against the rationality of salience-driven decisionmaking. There would be an argument against the rationality of salience-driven decisionmaking if discount retailers regularly made quality rather than price salient. But that is not the world that we live in.[12]

If we wanted to argue that salience-driven decisionmaking processes are not rational when vulnerable to framing effects, we would need to make one of three arguments. First, we could argue that salience-driven processes are insufficiently agential to warrant rational assessment:

> **(Arationality of Salience-Driven Decisionmaking)** Processes of salience-driven decisionmaking are neither rational nor irrational.

Arationality of Salience-Driven Decisionmaking would give us part of what we want, by

[11] Squirrel!

[12] This argument is naturally situated within the perspective of ecological rationality discussed in Section 7.

showing how framing effects can arise from processes that are not irrational, even if they are not rational either.

At the same time, an increasing number of philosophers argue that salience is a norm-governed matter.[13] Some theorists argue that harmful patterns of salience can constitute wrongs, as when an athlete's identity as a black woman or a rape victim is made more salient than her athletic achievements (Whiteley 2022, 2023). Others hold that prejudice itself constitutes in the misattribution of salience to properties such as race and gender (Munton 2023), so that obligations to avoid prejudice (Begby 2021; Basu 2019) will impose requirements on salience. Still others argue that democratic citizenship requires helping to make salient some facts of relevance to democratic life (Siegel 2022). If this is right, then Arationality of Salience-Driven Decisionmaking is too strong.

Second, we could claim that:

> **(Irrationality of Salience-Driven Decisionmaking)** Salience-driven decisionmaking is always irrational.

Irrationality of Salience-Driven Decisionmaking confronts the Problem of Forgone Advantages of Salience:

> **(Problem of Forgone Advantages of Salience)** Forgoing salience-driven decisionmaking deprives agents of the cognitive advantages of salience.

We need effective means to quickly select relevant information and eliminate distracting stimuli (Knudsen 2007), guide inference and choice (Munton 2023) and shape priority structures among mental representations (Watzl 2017). These are important benefits, which explain the centrality of salience in human cognition. If salience is not to be relied upon, then we need to be told how its benefits are to be reaped. Of if these benefits are not to be reaped, we need to be told why immunity to framing effects outweighs the central cognitive benefits provided by salience.

[13] As for the mechanism underlying norms of salience, it has been argued that salience may be under evaluative (Archer 2022) or long-range (Whiteley 2023) control, or may be norm-evaluable in virtue of its role in cognitive prioritization (Watzl 2017, 2022).

Finally, we could claim that:

> **(Frame-Independence of Rational Salience-Driven Decisionmaking)** Rational processes of salience-driven decisionmaking should be immune from framing effects.

Frame-Independence of Rational Salience-Driven Decisionmaking confronts the Problem of Forgone Advantages of Salience-Drivers:

> **(Problem of Forgone Advantages of Salience-Drivers)** Rendering salience-driven decisionmaking immune from framing effects deprives agents of useful drivers of salience.

Many of the most common and useful drivers of salience, such as contrast, surprise, and prominence, can be manipulated. I can decrease contrast by adding intermediate 'decoy' options (Bordalo et al. 2022), increase surprise by adding decision-irrelevant surprising features to objects in order to attract attention, or manipulate prominence by changing how prices are displayed. It is not clear that there are many, if any useful drivers of salience which cannot be used to generate framing effects. If Frame-Independence of Rational Salience-Driven Decisionmaking is not to collapse into Irrationality of Salience-Driven Decisionmaking, we are owed detailed and concrete examples of how salience can be made immune from framing effects while allowing salience to be sensitive to useful features of stimuli rather than less-useful features which are immune to manipulation.

5.4 Taking stock

This section introduced salience alongside three mechanisms driving stimulus salience: contrast, surprise and prominence (Section 5.1). A case study of the role of prominence in salience-based decisionmaking showed how salience can lead to framing effects when prominence is manipulated (Section 5.2). Nevertheless, we saw that salience plays important roles in human cognition, as a result of which there are good reasons to take salience-driven decisionmaking to be at least arational, and on many views fully rational, even when it exhibits moderate vulnerability to framing effects (Section 5.3). If this is right, then salience-based choice is a third mechanism generating procedurally rational framing effects.

6 Discussion

This paper argued that framing effects can be procedurally rational. Drawing on the frame-based choice model of Yuval Salant and Ariel Rubinstein (2008), we examined the case for:

> **(Procedural Rationality of Framing Effects)** For some agent *S*, time *t*, deliberative process *P*, and attitude *A*:
>
> > **(Formation)** *S* uses *P* to form *A* at *t*.
> >
> > **(Rational process)** It is rationally permissible for *S* to use *P* at *t*.
> >
> > **(Framing)** *P* exhibits framing effects.

We considered three models of how framing effects be procedurally rational: as the category-dependence of category-based choice (Section 3), order-dependence of choice from lists (Section 4), and salience-dependence of choice by agents who use salience to ration limited space in working memory (Section 5).

The direct upshot of this discussion is to lend support to the Procedural Rationality of Framing Effects. Procedural Rationality of Framing Effects complements recent philosophical attempts to show how framing effects can be in an important sense fully rational (Bermúdez 2020; Dreisbach and Guevara 2019). More broadly, Procedural Rationality of Framing Effects contributes to a growing program of vindicatory epistemology (Thorstad 2024), which seeks to show how apparently irrational cognitions are in an important sense fully rational (Gigerenzer and Sturm 2012; Icard ms; Lieder and Griffiths 2020).

Vindicatory epistemology can be put to many uses, including resisting nudging policies (Grüne-Yanoff and Hertwig 2016; Hertwig and Grüne-Yanoff 2017), understanding the causes and moral demands of poverty (Morton 2017; Thorstad forthcoming), grounding norms of inquiry (Thorstad 2024, forthcoming), advancing descriptive models of human cognition (Icard ms; Gigerenzer and Sturm 2012), and pushing back against dual process theories of cognition (Gigerenzer 2011; Thorstad 2025). I want to close by discussing a further application to moral debunking.

7 Bias and debunking

We saw in Section 1 that experimental philosophers and cognitive scientists have used the vulnerability of intuitions to cognitive bias as an argument against relying on moral and philosophical intuitions. This section presents a leading debunking strategy, the Master Argument of Walter Sinnott-Armstrong (Section 7.1). I use the results of this paper to develop two challenges to the Master Argument: the Uncoupling Challenge (Section 7.2) and the Frequency Challenge (Section 7.3). Section 7.4 concludes.

7.1 The Master Argument

Walter Sinnott-Armstrong (2008) presents the following Master Argument against noninferential reliance on moral intuitions:

(1) If our moral intuitions are formed in circumstances where they are unreliable, and if we ought to know this, then our moral intuitions are not justified without inferential confirmation.

(2) If moral intuitions are subject to framing effects, then they are not reliable in those circumstances.

(3) Moral intuitions are subject to framing effects in many circumstances.

(4) We ought to know this – that is, (3).

(5) Therefore, our moral intuitions in those circumstances are not justified without inferential confirmation. (Sinnott-Armstrong 2008, p. 52)

Early critics of Sinnott-Armstrong (Shafer-Landau 2008; Tolhurst 2008) identified a critical ambiguity in premise (2).

As in Section 2, say that:

> Choice function c **exhibits framing effects** if there exists a nonempty set $A \subseteq X$ and frames $f, f' \in \mathcal{F}$ with $c(A, f) \neq c(A, f')$.

A weak version of (2) holds that:

> **(2-Weak)** If an agent's choices are produced by a process which exhibits framing effects, then those choices are unreliable.

But the notion of unreliability grounded by Weak Reading cannot be the same notion used in (1). Choices which sometimes exhibit framing effects can be highly reliable if framing effects are infrequent. Instead, Sinnott-Armstrong needs:

> **(2-Strong)** If an agent's choices are produced by a process which *frequently* exhibits framing effects, then those choices are unreliable.

And hence also needs:

> **(3-Strong)** The processes which produce moral intuitions frequently exhibit framing effects.

This paper considers two ways of responding to (2-Strong) and (3-Strong).

The *Uncoupling Challenge* denies (2-Strong), holding that the reliability of processes can come uncoupled from the frequency with which they exhibit framing effects. The *Frequency Challenge* denies (3-Strong), holding that the processes which produce moral intuitions do not frequently exhibit framing effects. I take each challenge in turn.

7.2 The Uncoupling Challenge

One lesson from this paper is that the frequency with which processes exhibit framing effects can come apart from their reliability. Section 4 derived the optimal process of list-based choice using an explicit model of the likelihood that each framed decision problem would arise. We saw that the expected-utility-maximizing choice process in this situation, k-phase sastisficing, exhibits framing effects. Moreover, we saw that k-phase satisficing does not claim the virtue of exhibiting fewer framing effects than simpler competitors. Quite the opposite: k-phase satisficers accept a heightened vulnerability to framing effects in order to learn from experience during decisionmaking.

As the Uncoupling Challenge suggests, this discussion reveals that frequency with which a process exhibits framing effects is an imperfect guide to its reliability. It is, of course, true that framing effects are one driver of unreliability, but they are by no means the only driver of unreliability. As such, unless we observe a very high degree of frame-vulnerability, we cannot infer directly from the fact that a process frequently exhibits framing effects to the claim that this process is unreliable.

7.3 The Frequency Challenge

The Frequency Challenge holds that the processes which produce moral intuitions do not frequently exhibit framing effects. The usual way of pursuing the Frequency Challenge is to consider a particular context where for some choice set A and frames f, f', some agents exhibit the framing effect $c(A, f) \neq c(A, f')$. The usual strategy is to note that while some experimental participants show this pattern, many do not, so that any particular framing effect applies only to some agents (Demaree-Cotton 2016). This strategy may succeed, though it faces challenges. Many experiments find choice reversals across frames in at least 10-20% of participants, and sometimes the findings are stronger than this (Demaree-Cotton 2016; Machery 2017).

An alternative way to pursue the Frequency Challenge is to grant that in principle most agents' choice functions exhibit framing effects, but to question the likelihood that these effects will be realized in practice. For a given choice set A, this strategy considers the probabilities $Pr(f)$ that the agent is likely to encounter a particular framing (A, f). We then define the choice probability $Pr(c(A) = a)$ of each alternative by summing over framed decision problems, so that:

$$Pr(c(A) = a) = \sum_{f \in \mathcal{F} : c(A,f) = a} Pr(f). \tag{4}$$

On this reading, a process frequently exhibits framing effects to the extent that no option a has sufficiently large choice probability.

To show that a process frequently exhibits framing effects in the sense of (4), it is not enough to show that there are some framed decision problems (A, f) and (A, f') in which agents' decisions are likely to reverse. We must also show that those frames are likely to arise.

A preliminary motivation for thinking that framing effects are often infrequent in sense (4) is the ecological rationality paradigm in cognitive science (Morton 2017; Schmidt 2019; Todd and Gigerenzer 2012). On this paradigm, agents learn and evolve to select strategies that usually perform well in the environments where they are proposed for use. Once you know how a process works, you can manufacture environments in which that strategy shows framing effects and other performance errors. But unless those environments occur frequently and without agents' notice, agents may still be likely to deploy choice rules in contexts where they avoid framing effects and other errors. And if those environments did occur frequently, we might well learn or evolve to detect them and deploy different choice rules.

More generally, our defense of Procedural Rationality of Framing Effects suggests that framing effects produced by the processes surveyed in Sections 3-5 should not be too frequent in sense (4). A crucial part of the argument for the rationality of these processes was the argument that they are reliable in the environments where they are proposed for use. For example, we saw that salience allows agents to efficiently bring potentially relevant information into working memory, and that categorization allows agents to learn and apply relevant generalizations. To say that these processes are reliable is to say that the totality of irrelevant influences must not too often lead them astray. If that is right, then framing effects, as but one of many irrelevant influences, should also only infrequently lead reliable processes astray.

Let us illustrate this strategy for pursuing the Frequency Challenge by looking at the examples of category-based and salience-driven choice. It is well-known that categories and salience can be manipulated. But we do not usually think that framing effects of either type frequently alter choice. It is true that under the counterfactual assumption that Toyota just released an unusually high-quality van, my choice would have been different under the two categorizations discussed in Section 3. But Toyota released no such van, and more generally it is plausible that I would have settled on a Toyota Corolla under either categorization. Similarly, it is true that retailers occasionally convince me to buy candy bars through prominence manipulations, but we also saw in Section 5 that more global prominence manipulations to quantities such as price and quality often work in favor of, rather than against agents' efforts to make good choices.

The upshot of this discussion is that many processes which exhibit framing effects do not frequently do so. As a result, to establish (3-Strong), it is not enough to give examples of framing effects throughout moral decisionmaking. We need to think in detail about the processes which agents use to make choices and to argue that these processes should show frequent framing effects in sense (4). This argument is likely to rely on the claim that moral decisionmaking is often not driven by rational processes such as the processes surveyed in Sections 3-5, but instead by emotions such as fear and disgust (Haidt 2001) coupled with post-hoc rationalization (Greene 2008).

These procedural claims are controversial (May 2018). More to the point, if it could be established that human decisions are frequently made on the basis of such flagrantly irrational processes, then there would be far easier routes to establishing unreliability than an appeal to framing effects. It may be true that emotions such as fear and disgust can be manipulated across frames, but that is hardly the strongest reason to believe that these emotions are often poor guides to choiceworthiness. If this is right, then once we shift our attention to questions about the particular processes responsible for framing effects, we may find ourselves less concerned with questions about framing effects and more concerned with settling broader debates about the kinds of processes driving moral decisionmaking.

7.4 Taking stock

This section explored the Master Argument of Sinnott-Armstrong (2008) against noninferential reliance on moral intuitions (Section 7.1). We saw that the findings of this paper support two strategies for resisting the Master Argument. The Uncoupling Challenge (Section 7.2) holds that the frequency with which a process exhibits framing effects can come uncoupled from its reliability, as in the example of *k*-phase satisficing from Section 4. The Frequency Challenge (Section 7.3) suggests that while many processes exhibit some vulnerability to framing effects, they may not frequently exhibit framing effects. This happens not only because many experimental participants fail to show any given framing effect, but also because the frames required to elicit framing effects may occur rarely outside of a laboratory context. And we saw that the best way of resisting these challenges may involve a diminished emphasis on framing effects as a guide to the reliability of moral intuitions.

References

Anderson, John. 1990. *The adaptive character of thought*. Lawrence Erlbaum Associates.

Archer, Sophie. 2022. "Salience and what matters." In Sophie Archer (ed.), *Salience: A philosophical inquiry*, 113–29. Routledge.

Banovetz, James and Oprea, Ryan. 2023. "Complexity and procedural choice." *American Economic Journal: Microeconomics* 15:384–413.

Basu, Rima. 2019. "Radical moral encroachment: The moral stakes of racist beliefs." *Philosophical Issues* 29:9–23.

Begby, Endre. 2021. *Prejudice: A study in non-ideal epistemology*. Oxford University Press.

Bermúdez, José. 2020. *Frame it again*. Cambridge University Press.

—. 2022. "Rational framing effects: A multidisciplinary case." *Behavioral and Brain Sciences* 45:e220.

Binz, Marcel and Schulz, Eric. 2023. "Using cognitive psychology to understand GPT-3." *Proceedings of the National Academy of Sciences* 120:e2218523120.

Bjork, Robert A and Whitten, William B. 1974. "Recency-sensitive retrieval processes in long-term free recall." *Cognitive Psychology* 6:173–89.

Bordalo, Pedro, Gennaioli, Nicola, and Shleifer, Andrei. 2012. "Salience theory of choice under risk." *Quarterly Journal of Economics* 127:1243–85.

—. 2016. "Competition for attention." *Review of Economic Studies* 83:481–513.

—. 2020. "Memory, attention and choice." *Quarterly Journal of Economics* 135:1399–442.

—. 2022. "Salience." *Annual Review of Economics* 14:521–44.

Browne, Katharine and Watzl, Sebastian. 2025. "The attention market - and what is wrong with it." *Philosophical Studies* https://doi.org/10.1007/s11098--025--02436--3.

Castro, Clinton and Pham, Adam. 2020. "Is the attention economy noxious?" *Philosophers' Imprint* 20:1–13.

Chwang, Eric. 2016. "Consent's been framed: When framing effects invalidate consent and how to validate it again." *Journal of Applied Philosophy* 33:270–85.

Cohen, Shlomo. 2013. "Nudging and informed consent." *American Journal of Bioethics* 13:3–11.

Demaree-Cotton, Joanna. 2016. "Do framing effects make moral intuitions unreliable?" *Philosophical Psychology* 29:1–22.

Director, Samuel. 2024. "Framing effects do not undermine consent." *Ethical Theory and Moral Practice* 27:221–35.

Dreisbach, Sandra and Guevara, Daniel. 2019. "The Asian disease problem and the ethical implications of prospect theory." *Noûs* 53:613–38.

Fischoff, Baruch. 1982. "Debiasing." In Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), *Judgment under uncertainty: Heuristics and biases*, 422–44. Cambridge University Press.

Gabaix, Xavier. 2014. "A sparsity-based model of bounded rationality." *Journal of Economic Literature* 1661–1710.

Geman, Stuart, Bienenstock, Elie, and Doursat, René. 1992. "Neural networks and the bias/variance dilemma." *Neural Computation* 4:1–58.

Gerken, Mikkel. 2022. "Salient alternatives and epistemic injustice in folk epistemology." In Sophie Archer (ed.), *Salience: A philosophical inquiry*, 213–34. Routledge.

Gigerenzer, Gerd. 2011. "Personal reflections on theory and psychology." *Theory and Psychology* 20:733–43.

—. 2018. "The bias bias in behavioral economics." *Review of Behavioral Economics* 5:303–336.

Gigerenzer, Gerd and Selten, Reinhard (eds.). 2001. *Bounded rationality: The adaptive toolbox*. MIT Press.

Gigerenzer, Gerd and Sturm, Thomas. 2012. "How (far) can rationality be naturalized?" *Synthese* 187:243–68.

Greene, Joshua. 2008. "The secret joke of Kant's soul." *Moral Psychology* 3:35–79.

Grüne-Yanoff, Till and Hertwig, Ralph. 2016. "Nudge versus boost: How coherent are policy and theory?" *Minds and Machines* 26:149–83.

Guney, Begum. 2014. "A theory of iterative choice in lists." *Journal of Mathematical Economics* 53:26–32.

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: A social intuitionist approach to moral judgment." *Psychological Review* 108:814–34.

Hanna, Jason. 2011. "Consent and the problem of framing effects." *Ethical Theory and Moral Practice* 14:517–31.

Hertwig, Ralph and Grüne-Yanoff, Till. 2017. "Nudging and boosting: Steering or empowering good decisions." *Perspectives on Psychological Science* 12:973–86.

Horne, Zachary and Livengood, Jonathan. 2015. "Ordering effects, updating effects, and the specter of global skepticism." *Synthese* 194:1189–1218.

Horowitz, Tamara. 1998. "Philosophical intuitions and psychological theory." *Ethics* 108:367–85.

Horvath, Joachim. 2010. "How (not) to react to experimental philosophy." *Philosophical Psychology* 23:447–80.

Horvath, Joachim and Wiegmann, Alex. 2020. "Intuitive expertise in moral judgments." *Australasian Journal of Philosophy* 100:342–59.

Icard, Thomas. ms. *Resource rationality*.

Jones, Erik and Steinhardt, Jacob. 2022. "Capturing failures of large language models via human cognitive biases." In *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 11785–99.

Kamm, Francis. 1998. "Moral intuitions, cognitive psychology, and the harming-versus-not-aiding distinction." *Ethics* 108:463–88.

Knudsen, Eric. 2007. "Fundamental components of attention." *Annual Review of Neuroscience* 30:57–78.

Larrick, Richard. 2004. "Debiasing." In Derek Koehler and Nigel Harey (eds.), *Blackwell handbook of judgment and decisionmaking*, 316–38. Blackwell.

Levin, Irwin and Gaeth, Gary. 1988. "How consumers are affected by the framing of attribute information before and after consuming the product." *Journal of Consumer Research* 15:374–8.

Levin, Irwin, Schneider, Sandra, and Gaeth, Gary. 1998. "All frames are not created equal: A typology and critical analysis of framing effects." *Organizational Behavior and Human Decision Processes* 76:149–88.

Li, Ye and Epley, Nicholas. 2009. "When the best appears to be saved for last: Serial position effects on choice." *Journal of Behavioral Decision Making* 22:378–89.

Lieder, Falk and Griffiths, Thomas. 2020. "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources." *Behavioral and Brain Sciences* 43:E1.

Machery, Edouard. 2017. *Philosophy within its proper bounds*. Oxford University Press.

Mantonakis, Antonia, Rodero, Pauline, Lesschaeve, Isabelle, and Hastie, Reid. 2009. "Order in choice: Effects of serial position on preferences." *Psychological Science* 20:1309–12.

Manzini, Paola and Mariotti, Marco. 2012. "Categorize then choose: Boundedly rational choice and welfare." *Journal of the European Economic Association* 10:1141–65.

May, Joshua. 2018. *Regard for reason in the moral mind*. Oxford University Press.

Medin, Douglas and Heit, Evan. 1999. “Categorization.” In Benjamin Bly and David Rumelhart (eds.), *Cognitive science*, 99–143. Academic Press.

Miller, Jaonne and Krosnick, Jon. 1998. “The impact of candidate name order on election outcomes.” *Public Opinion Quarterly* 62:291–330.

Mohlin, Erik. 2014. “Optimal categorization.” *Journal of Economic Theory* 152:365–81.

Morton, Jennifer. 2017. “Reasoning under scarcity.” *Australasian Journal of Philosophy* 95:543–59.

Munton, Jessie. 2023. “Prejudice as the misattribution of salience.” *Analytic Philosophy* 64:1–19.

Rabin, Michael and Scott, Dana. 1959. “Finite automata and their decision problems.” *IBM Journal of Research and Development* 3:114–25.

Rubinstein, Ariel. 1986. “Finite automata play repeated prisoner’s dilemma.” *Journal of Economic Theory* 39:83–96.

Rubinstein, Ariel and Salant, Yuval. 2006. “A model of choice from lists.” *Theoretical Economics* 1:3–17.

Salant, Yuval. 2011. “Procedural analysis of choice rules with applications to bounded rationality.” *American Economic Review* 101:724–48.

Salant, Yuval and Rubinstein, Ariel. 2008. “(A,f): Choice with frames.” *Review of Economic Studies* 75:1287–96.

Sandborn, Adam, Griffiths, Thomas, and Navarro, Daniel. 2006. “A more rational model of categorization.” In *Proceedings of the 28th annual conference of the Cognitive Science Society*, 726–31. Lawrence Erlbaum.

Schmidt, Andreas. 2019. “Getting real on rationality – behavioral science, nudging, and public policy.” *Ethics* 129:511–543.

Schwitzgebel, Eric and Cushman, Fiery. 2012. "Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers." *Mind and Language* 27:135–53.

Shafer-Landau, Russ. 2008. "Defending ethical intuitionism." In Walter Sinnott-Armstrong (ed.), *Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity*, 83–95. MIT Press.

Siegel, Susanna. 2022. "Salience principles for democracy." In Sophie Archer (ed.), *Salience: A philosophical inquiry*, 235–66. Routledge.

Simon, Herbert. 1971. "Designing organizations for an information-rich world." In Martin Greenberger (ed.), *Computers, communications, and the public interest*, 37–72. Johns Hopkins Press.

—. 1976. "From substantive to procedural rationality." In T.J. Kastelein, S.K. Kuipers, W.A. Nijenhuls, and R.G. Wagenaar (eds.), *25 years of economic theory*, 65–86. Springer.

Sims, Christopher. 2003. "Implications of rational inattention." *Journal of Monetary Economics* 50:665–90.

—. 2006. "Rational inattention: Beyond the linear-quadratic case." *American Economic Review* 96:158–63.

Sinnott-Armstrong, Walter. 2008. "Framing moral intuitions." In Walter Sinnott-Armstrong (ed.), *Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity*, 47–76. MIT Press.

Stein, Edward. 1996. *Without good reason: The rationality debate in philosophy and cognitive science*. Clarendon Press.

Sunstein, Cass. 2014. *Why nudge? The politics of libertarian paternalism*. Yale University Press.

Swain, Stacey, Alexander, Joshua, and Weinberg, Jonathan. 2008. "The instability of philosophical intuitions: Running hot and cold on truetemp." *Philosophy and Phenomenological Research* 76:138–55.

Thaler, Richard and Sunstein, Cass. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Thorstad, David. 2024. *Inquiry under bounds*. Oxford University Press.

—. 2025. "The complexity-coherence tradeoff in cognition." *Mind* 134:422–57.

—. forthcoming. "The zetetic turn and the procedural turn." *Journal of Philosophy* forthcoming.

Todd, Peter and Gigerenzer, Gerd. 2012. *Ecological rationality: Intelligence in the world*. Oxford University Press.

Tolhurst, William. 2008. "Moral intuitions framed." In Walter Sinnott-Armstrong (ed.), *Moral psychology, volume 2: The cognitive science of morality: Intuition and diversity*, 77–82. MIT Press.

Tversky, Amos and Kahneman, Daniel. 1981. "The framing of decisions and the psychology of choice." *Science* 211:453–8.

Viale, Riccardo. 2021. "Why bounded rationality?" In Riccardo Viale (ed.), *Routledge handbook of bounded rationality*, 1–54. Routledge.

Watzl, Sebastian. 2017. *Structuring the mind: The nature of attention and how it shapes consciousness*. Oxford University Press.

—. 2022. "The ethics of attention: An argument and a framework." In Sophie Archer (ed.), *Salience: A philosophical inquiry*, 89–112. Routledge.

Weinberg, Jonathan, Gonnerman, Chad, Buckner, Cameron, and Alexander, Joshua. 2010. "Are philosophers expert intuiters?" *Philosophical Psychology* 23:331–55.

Whiteley, Ella. 2022. "Harmful salience perspectives." In Sophie Archer (ed.), *Salience: A philosophical inquiry*, 193–212. Routledge.

Whiteley, Ella Kate. 2023. "'A woman first and a philosopher second': Relative attentional surplus on the wrong property." *Ethics* 133:497–528.

Wiegmann, Alex, Okan, Yasmina, and Nagel, Jonas. 2012. "Order effects in moral judgment." *Philosophical Psychology* 25:813–36.

Williamson, Timothy. 2011. "Philosophical expertise and the burden of proof." *Metaphilosophy* 42:215–29.