PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Computational mechanisms for context-based behavioral interventions: A large-scale analysis

Wenjia Joyce Zhao[a,b,1], Aoife Coady[b], and Sudeep Bhatia[b,c]

Edited by Richard Shiffrin, Indiana University Bloomington, Bloomington, IN; received August 12, 2021; accepted February 18, 2022

Choice context influences decision processes and is one of the primary determinants of what people choose. This insight has been used by academics and practitioners to study decision biases and to design behavioral interventions to influence and improve choices. We analyzed the effects of context-based behavioral interventions on the computational mechanisms underlying decision-making. We collected data from two large laboratory studies involving 19 prominent behavioral interventions, and we modeled the influence of each intervention using a leading computational model of choice in psychology and neuroscience. This allowed us to parametrize the biases induced by each intervention, to interpret these biases in terms of underlying decision mechanisms and their properties, to quantify similarities between interventions, and to predict how different interventions alter key choice outcomes. In doing so, we offer researchers and practitioners a theoretically principled approach to understanding and manipulating choice context in decision-making.

decision-making | behavioral interventions | context effects | computational modeling

Choices depend on incidental contextual factors, such as defaults, social norms, and time pressure, and understanding how these contextual factors influence choice has been a major focus of research in the behavioral sciences (1–3). Context dependence also forms the psychological rationale behind behavioral interventions, or "nudges" (4). Such interventions influence behavior by altering choice context and are important tools for policy makers, as they can improve decisions without restricting people's capacity to choose.

Although useful qualitative taxonomies of context effects and behavioral interventions have been proposed, none offers a cognitively and neurobiologically inspired model [such as in refs. (5, 6)] with which to parametrize and predict the effect of many different contextual factors on behavior. There are two major challenges to accomplishing this goal. The first is the lack of data: Large-scale experiments that collect data on multiple behavioral interventions within a single decision task are necessary to build unified models, but these experiments are very rare. The second involves the psychological complexity of context, which can alter the choice predispositions that decision makers have prior to evaluating the choice options, how desirable the options appear to the decision makers during evaluation, as well as how hard decision makers are willing to deliberate. It is not clear how theorists should incorporate these diverse effects into a single computational model of choice.

In this article, we address these challenges using a comprehensive data set of choices under the influence of 19 contextual factors. These contextual factors include a diverse set of situational variables that have been shown to influence choice without altering people's capacity to choose (for simplicity, we do not examine the role of explicit persuasion, financial incentivization, direct education, or coercion, which are also important determinants of choice). In our experiments, we assigned more than 1,200 laboratory participants to different context-based behavioral interventions (adapted from existing research) in two decision scenarios involving either consumer or financial choice. Participants made 160 binary choices both with and without the interventions, resulting in more than 300,000 total choices and response times (RTs; i.e., the time taken to make decisions) for quantitative analysis. The interventions are summarized in Table 1, and an example of the experimental stimuli in experiment (exp.) 1 is shown in Fig. 1*A* (exp. 2 involved nearly identical stimuli presentation). Additional details are provided in *Methods* and experimental instructions are provided in *SI Appendix*, Tables S1 and S2.

We modeled our choice data set with the diffusion decision model (DDM), which proposes that decision makers dynamically accumulate the relative evidence favoring each choice option. Choices are made when the relative evidence reaches one of two decision boundaries, and the time to reach the decision boundary corresponds to the RT. DDM is the dominant computational model of two-alternative forced choice in psychology and neuroscience (7–9) and has been shown to successfully describe binary

Significance

A large body of research in the social and behavioral sciences studies the impact of behavioral interventions (or "nudges") on decisions. Although this work has been extremely influential, we currently lack an overarching theoretical framework for behavioral interventions that provides a systematic account of their behavioral consequences, cognitive and neurobiological mechanisms, and statistical interpretations. In this paper, we propose such a theoretical framework using the diffusion decision model, a quantitative theory of decision-making whose parameters offer a theoretically compelling characterization of choice underpinnings. Our results not only reveal insights about how context-based interventions alter behavior but also offer practitioners a model-based method for choosing between behavioral interventions based on different goals.

Author affiliations: [a]Department of Psychology, The Ohio State University, Columbus, OH 43210; [b]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; and [c]Department of Marketing, University of Pennsylvania, Philadelphia, PA 19104

Author contributions: W.J.Z., A.C., and S.B. designed research; W.J.Z. and A.C. performed research; W.J.Z. and S.B. analyzed data; and W.J.Z. and S.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: joyce.wenjia.zhao@gmail.com.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2114914119/-/DCSupplemental.

Published April 4, 2022.

Table 1. Summary of behavioral interventions

Category	Intervention	Procedure and instruction
Prominence of Information	1. Attribute order	Within-participant manipulation. The quality attribute was positioned above or below the price attribute.
	2. Option order	Within-participant manipulation. The high-quality option was positioned to the left or right of the low-price option.
	3. Quality priming	Implemented in exp. 1 only. Photos of appetizing and expensive food were shown on each instruction page (before the task and between blocks).
	4. Price priming	Implemented in exp. 1 only. Photos of US dollars were shown on each instruction page (before the task and between blocks).
	5. Quality information	Implemented in exp. 2 only. Participants read a short passage explaining health insurance deductibles, and they answered multiple choice questions about why a low-deductible health insurance plan could be beneficial.
	6. Price information	Implemented in exp. 2 only. Participants read a short passage explaining health insurance premiums, and they answered multiple choice questions about why a low-premium health insurance plan could be beneficial.
	7. Attribute salience	One of the two attributes appeared with an orange frame, which highly contrasted with the background.
	8. Option salience	One of the two options appeared with an orange frame, which highly contrasted with the background.
Task framing	9. Default	One of the options was preselected, and additional key pressing was required to switch the option.
	10. Reject (vs. accept)	Participants indicated which option they preferred less instead of indicating which option they preferred more.
Social information	11. Social norm	The more popular option in each choice problem (based on a pilot study) was indicated using an orange frame.
	12. Recommendation	The option recommended by the experimenters (based on the same pilot study for condition 11) was indicated using an orange frame.
Affect	13. Positive emotion	Before the choice task, participants took 5–10 min to write a report of a happy event from their life. They were also instructed to reread the event during each break.
	14. Negative emotion	Before the choice task, participants took 5–10 min to write a report of a sad event from their life. They were also instructed to reread the event during each break.
Speed and accuracy	15. Time pressure	Participants were instructed to make choices as quickly as possible.
	16. Accuracy instruction	Participants were asked to indicate choices only after they were completely certain about their choice.
	17. Cognitive load	Participants performed an additional memory task in which they remembered a six-digit number before each block and reported the number at the end of each block.
	18. Accountability	Participants wrote a one-paragraph justification of one of their choices (randomly selected at the end of the choice task).
	19. Font fluency	The stimuli were shown in a hard-to-read font.

Note: All interventions that are not explicitly listed as "within-participant" are "between-participant."

responses in perception, memory, categorization, language, and value-based decision-making. We fit a hierarchical DDM to all participants assigned to a given behavioral intervention condition ($n \sim 40$ participants in each condition, with each participant completing 160 baseline trials and 160 intervention trials; details are provided in *Methods*). For each fit, we quantified the effect of the behavioral intervention in terms of how it shifted three key DDM parameters: the starting point, the drift rate, and the decision boundary. Changes in these three parameters have systematic effects on choice probability and RT. All else being equal, starting points favoring one option over the other generate more frequent choices and quicker responses for the favored option; higher drift rates favoring one option over the other generate more frequent choices for the favored option, with higher absolute drift rates causing quicker responses for both options; and higher decision boundaries lead to slower responses and more consistent choices. Fig. 1*B* summarizes the DDM and illustrates these predictions in a hypothetical choice between a high-quality and a low-price option. Although the choice probability for the high-quality option increases in all three panels, RT distributions vary based on the parameter that is causing the shift. For this reason, changes to DDM parameters can be identified based on choice and RT data.

The DDM allows us to model many different behavioral interventions within a single computational framework. Importantly, it also allows us to interpret their effects in terms of underlying cognitive and neurobiological mechanisms, as well as their statistical properties (10–12). The starting-point

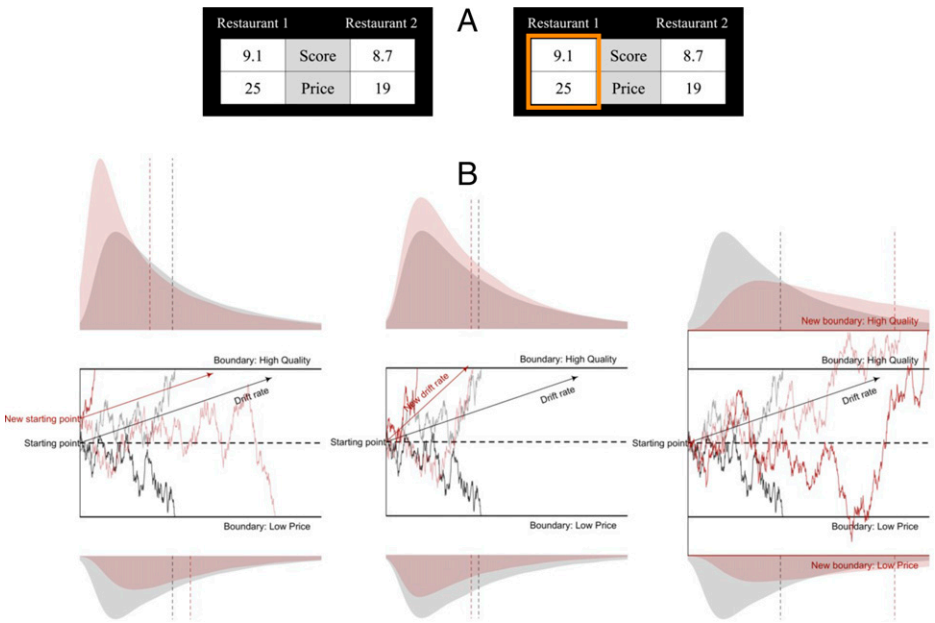

Fig. 1. (*A*) Screenshots for exp.1, with an example of the baseline condition (*Left*) and an option salience intervention (*Right*). (*B*) Illustration of the DDM, with hypothetical changes to the starting point, drift rate and decision boundary parameters.

parameter, for example, corresponds to an automatic predecisional bias (or, equivalently, a baseline activation bias in neural units) and can be interpreted as a statistical prior regarding the relative desirability of the available options. The drift rate governs the formation of preferences (or, equivalently, the rate of change of neural activation) during deliberation and reflects evolving probabilistic beliefs as decision makers integrate evidence favoring one option over the other. Last, the decision boundary is a threshold level of preference (or activation) strength necessary to initiate choice and can be seen as specifying an acceptable error rate for the decision. By observing how different contextual factors influence these three DDM mechanisms, we are able to study the behavioral, cognitive, neural, and statistical consequences of decision context on choice.

Results

Overall Trends. The decision scenarios in exp. 1 and exp. 2 were taken from consumer and financial choice domains, respectively. The former offered participants choices between pairs of restaurants (each with a user rating and a price); the latter offered participants choices between pairs of health insurance plans (each with a deductible and a premium). These two scenarios were tested in two separate preregistered experiments (Open Science Framework, https://osf.io/y59jk/). Each experiment involved 15 between-participant interventions. Thirteen between-participant interventions were implemented in both experiments. Money priming and food-quality priming were implemented only in exp. 1, and informational prompts emphasizing the importance of either low deductibles or low premiums were implemented only in exp. 2. Finally, in addition to the 17 between-participant interventions across the two experiments, we also counterbalanced the presentation order of choice options and attributes within participants in each experiment. This resulted in a total of 19 context-based interventions for our analysis. Note that for expositional simplicity, we refer to high-rating/high-price restaurants in exp. 1 and low-deductible/high-premium insurance plans in exp. 2 as "high-quality" options. Low-rating/low-price restaurants and high-deductible/low-premium insurance plans are referred to as "low-price" options.

Overall, choices in the experiments were fairly balanced, with 53.5% and 56.8% of choices favoring the high-quality options in exp. 1 and exp. 2, respectively. As expected, different behavioral interventions had different choice probability and RT effects (*SI Appendix*, Fig. S1). Our DDMs captured this variability, with correlations greater than 0.94 in predicting individual-level changes to choice probabilities and RTs, and correlations greater than 0.97 in predicting aggregate changes to choice probabilities and RTs, across interventions. These patterns are shown in Fig. 2.

Our main goal was to quantify and compare the effects of the interventions on the three DDM parameters. These parameter effects are summarized in Fig. 3, which displays the shift in the starting point (Fig. 3*A*), the drift rate (Fig. 3*B*), and the decision boundary (Fig. 3*C*) for each intervention and each experiment. In the following paragraphs of *Results*, we analyze the effects of social norms, recommendations, defaults, and option and attribute salience separately for trials selectively targeting the high-quality and low-price options in the two experiments). *SI Appendix*, Fig. S2 provides means and 95% CIs of DDM parameter shifts, *SI Appendix*, Fig. S3 shows that the behavioral effects of interventions are correlated with the estimated effects of interventions on DDM parameters, *SI Appendix*, Figs. S4 and S5 provide results of cross-validation predictions of choice probabilities and RT percentiles, *SI Appendix*, Fig. S6 shows that the DDM parameters are recoverable, and *SI Appendix*, Fig. S9 shows that DDM parameters correlate strongly with those estimated using an alternate evidence accumulation model.

Starting Point Effects. In Fig. 3*A*, we see that the interventions with the strongest effect on starting points in the two experiments were the social norm (13, 14) and recommendation (15) interventions. This indicates that participants developed an automatic predecisional response bias (equivalently, a baseline activation bias or statistical prior) in favor of the norm or recommendation, increasing its choice probability and reducing its RT. These results are consistent with prior work in which authors found that social cues can alter the automatic processing of choice options (16).

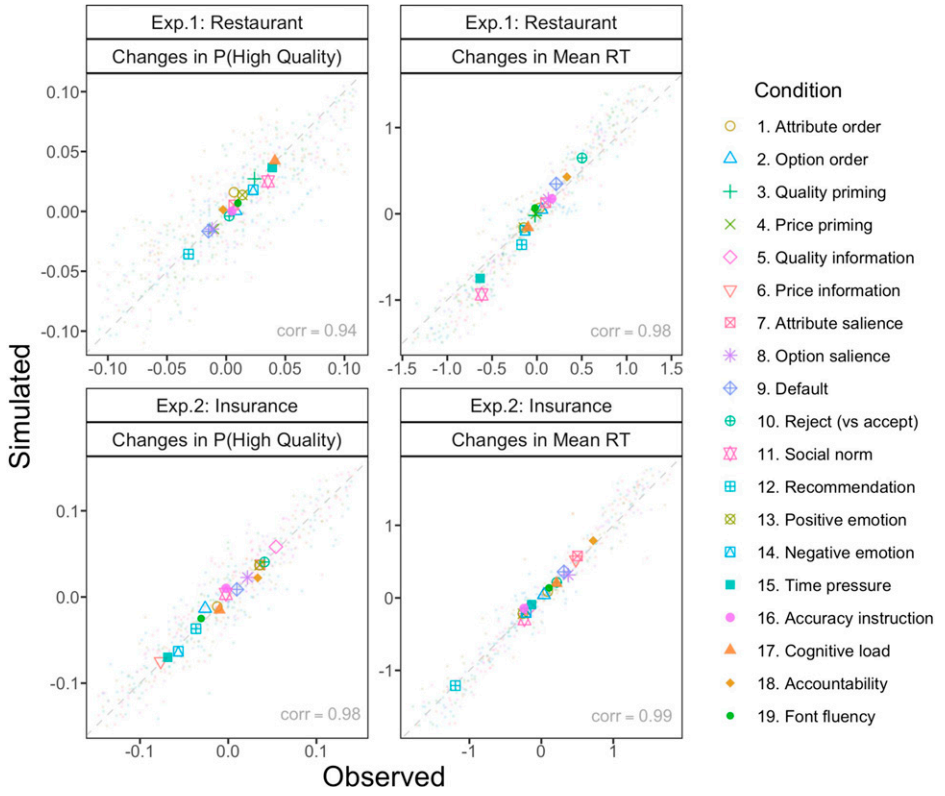

Fig. 2. Observed and simulated intervention effects on choice probabilities and RTs in exp.1 and exp. 2. Gray points correspond to changes for individual participants. Colored labels correspond to aggregate changes for interventions, averaged over participants. Displayed correlations capture the relationship between observed and simulated changes on the individual level. Note: Participants with behavioral shifts outside of the range of the *x* and *y* axes of this figure are shown in *SI Appendix*, Fig. S11.

Defaults (17, 18) also had positive effects on starting points in both experiments, so that making an option a default typically generated a mild predecisional bias for that option. We obtained similar but weaker results for option salience (19, 20). These findings validate previous claims regarding the automatic effects of defaults and salient options (21, 22). Additionally, informational prompts about the importance of low premiums (low price) biased starting points in favor of the corresponding health insurance plans in exp. 2. We did not observe the analogous effect for informational prompts about the importance of low deductibles (high quality).

Overall, the starting point effects of the interventions were largely consistent across the two experiments, with a Spearman correlation of 0.46 for the interventions. That said, some interventions had stronger effects in one experiment than the other. For example, participants were more likely to display a biased starting point in favor of the low-price option when asked to reject (instead of accepting) (23) one of the options in exp. 1. We did not find a systematic starting point effect for reject framing in exp. 2, in which both attributes were monetary. This could reflect a tendency to prioritize monetary attributes in rejection decisions. Conversely, negative emotion (24, 25) generated a starting point bias in exp. 2 (favoring the high-quality insurance plan) but did not influence starting points in exp. 1. Positive emotions did not yield analogous effects, suggesting that these effects could have involved negative emotions like fear and sadness, which are known to bias choices in favor of the safer choice (26). We also found a difference between the two domains for font fluency (27): Disfluent fonts led to a weak bias favoring the low-price insurance plan in exp. 2 but had no such effect in exp. 1. We speculate that this could reflect differences in the importance or difficulty of the two decisions. Finally, we note that some starting point effects could be a product of unique features of our experimental design, such as repeated choices. Experience leads to priors (and subsequently starting point effects) favoring preferred options, and inexperienced individuals in the real-world making one-off choices may not display such biases.

Drift Rate Effects. Unsurprisingly, we found that informational prompts had an effect on the drift rates in exp. 2 (favoring the option that was dominant on the prompted attribute; Fig. 3*B*). Thus, informational prompts alter the evaluation of choice options. Likewise, priming money or quality (28) led to corresponding drift rate effects in exp. 1.

Drift rates were also very strongly influenced by social norm and experimenter recommendation interventions in both experiments. Thus, in addition to developing predecisional biases favoring normative and recommended options, participants also developed explicit preferences for these options. This finding is consistent with prior work reporting that social cues shift people's preferences in favor of conforming options (16). By contrast, default and salient options did not have drift rate effects. Decision makers did begin with predecisional biases for these options but did not display biased preferences for these options once they began deliberating. The null effects of defaults are consistent with prior work suggesting that defaults operate primarily through automatic and nonevaluative mechanisms (4). The null effects of salience are consistent with research suggesting the disproportionate effect of salience and accessibility on automatic (vs. deliberative) processing (22).

Again, drift rate effects were consistent across the two experiments, with a Spearman correlation of 0.39. However, we found that cognitive load (29) and time pressure (30) biased

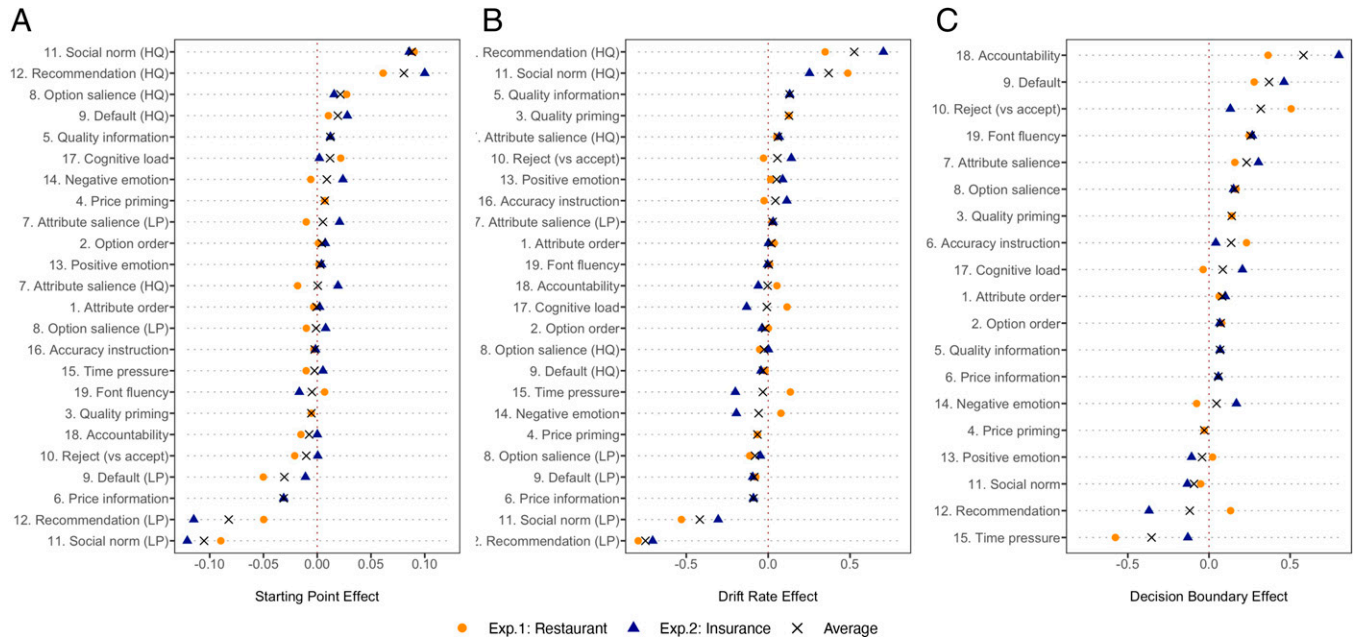

Fig. 3. Effects of behavioral interventions on (*A*) the start point, (*B*) the drift rate, and (*C*) the decision boundary. Positive (negative) starting point and drift rate effects correspond to biases favoring the high-quality (low-price) option. HQ and LP denote interventions selectively targeting the high-quality or low-price options in a trial. Results are based on group-level parameters in each condition.

exp. 1 drift rates in favor of the high-quality option, but exp. 2 drift rates were biased in favor of the low-price option. It seems, therefore, that such interventions do not only alter the amount of time and effort participants are willing to put into the decision but also how they evaluate the choice options, specifically, their preferences for price over quality, which may vary across decision scenarios. This may reflect an adaptive response to effort or accuracy tradeoffs in decision-making (31, 32). Finally, as with the starting point effects discussed previously, we found that negative emotion had a stronger effect on drift rates in exp. 2 than in exp. 1 and that the effect of positive emotion was almost completely absent in the two experiments. This indicates that the emotion effects we documented persist beyond predecisional bias and influence the evaluation of options and the formation of preferences.

Decision Boundary Effects. The final mechanism we analyzed was the decision boundary, whose effects are shown in Fig. 3*C*. Here, we can see that increasing decision-maker accountability (33) had a strong positive effect on the boundary for both experiments. Intuitively, decision makers who were accountable for their decisions deliberated longer and required a greater degree of confidence (or equivalently, activation) before choosing. This is equivalent to having a lower acceptable error probability for the decision. Our finding is consistent with that of considerable prior work on accountability in decision-making (33). We also found a positive effect of font-fluency interventions on the decision boundary in both experiments: Making the font disfluent has also been found to increase deliberative processing (27).

Other interventions that had a strong effect on the decision boundary were reject (vs. accept) framing and defaults, which increased the boundaries in both experiments. Both interventions implicitly involved rejection decision frames, which have been shown to increase RTs when options are desirable (34). Additionally, as expected, time pressure had a strong negative effect on the decision boundary. When asked to make quicker decisions, decision makers reduced their thresholds and increased acceptable error levels.

Once again, the decision boundary effect was consistent across the two conditions, with a Spearman correlation of 0.51. However, interestingly, recommendations reduced decision boundaries in exp. 2 but not in exp. 1, suggesting that participants needed less evidence before deciding when given social information in more important or more difficult decisions. We also found a positive effect of accuracy instructions on decision boundary for exp. 1 but not for exp. 2, possibly because exp. 2 already had high boundaries.

Comparative Effects. Our analytical approach allows us to describe distinct contextual effects within a unified representational framework. This can shed light on the similarities and differences between different behavioral interventions. We illustrate this in Fig. 4. Fig. 4*A* plots each intervention in each of our two experiments in a single three-dimensional space of behavioral interventions. The coordinates for each intervention capture its absolute, standardized effects on the starting point, drift rate, and decision boundary parameters. Here, we can see that the different behavioral interventions span a wide region of the space, with some interventions having similar effects on all three mechanisms and others having a stronger effect on one or two of the mechanisms.

Fig. 4*B* displays the overall magnitude of the intervention vectors. Vector magnitude captures the total cognitive effect size of the interventions, that is, the cumulative absolute effect of the intervention on the drift rates, starting points, and decision boundaries of our DDM models. This figure shows that recommendations, social norms, accountability, time pressure, and defaults had the strongest effects across decision scenarios, whereas emotions, priming, and accuracy instructions had the weakest effects across our decision scenarios.

Finally, interventions also had different degrees of similarity to each other, as revealed by the relative positions of the vectors in Fig. 4*A*. We show this more rigorously by clustering the interventions in Fig. 4*C*. Here, we see that the social norm and

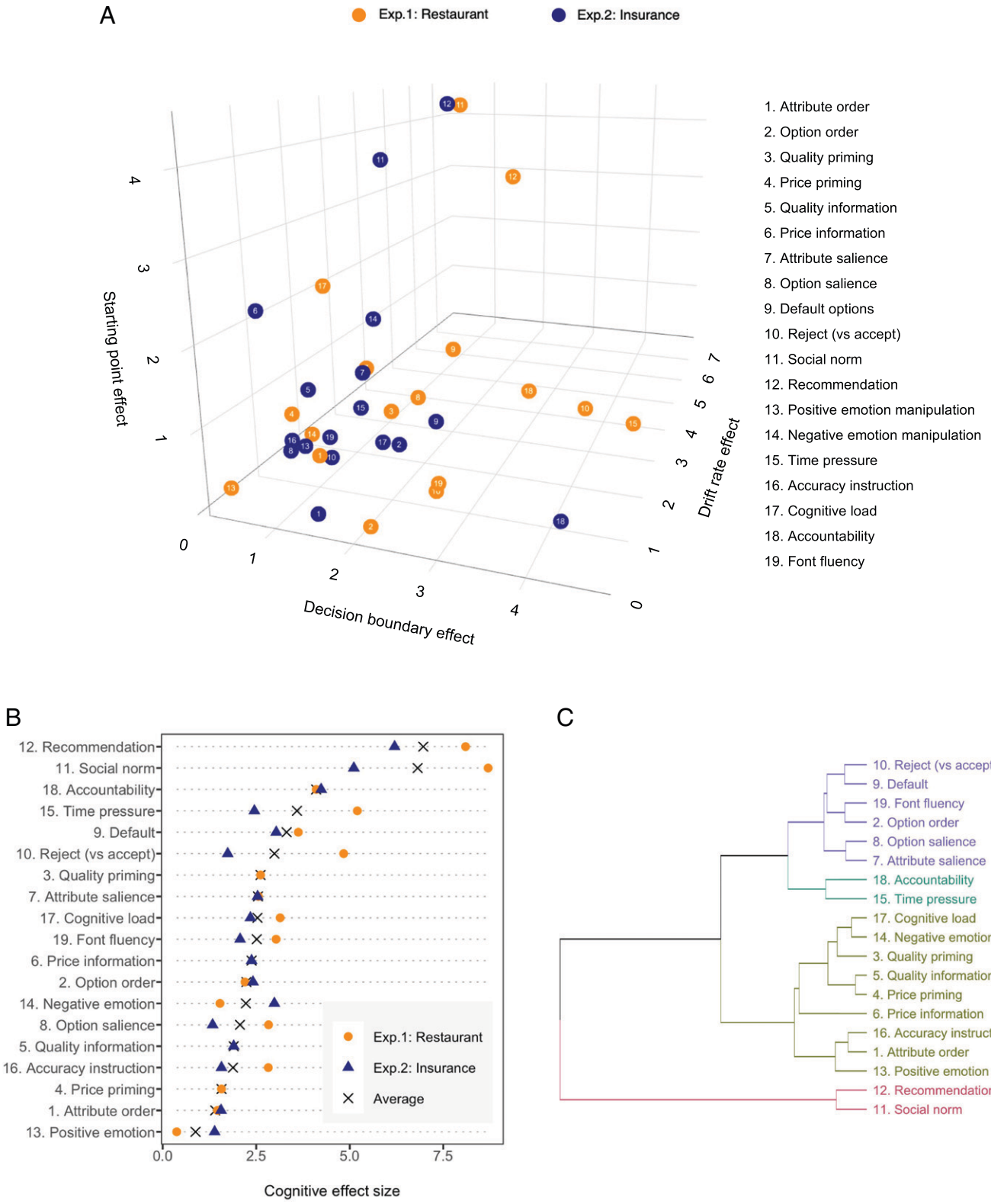

Fig. 4. (*A*) Three-dimensional space of behavioral interventions based on their absolute, standardized effects on the starting point, drift rate, and decision boundary parameters. (*B*) Cognitive effect sizes of the interventions. These are based on the distance between an intervention and the origin of the space. (*C*) Hierarchical clustering of intervention vectors (averaged across exp.1 and exp. 2). Results are based on group-level parameters in each condition.

recommendation interventions cluster together. These typically involve social cues that bias the decision maker in favor of one option over the other. We also found that default and rejection interventions (which involve the manipulation of task frames) clustered together, as did salience manipulations (which involve changes to information prominence). Finally, accountability and time pressure manipulations shared a cluster, as they both involved altering accuracy and speed tradeoffs.

Discussion

Computational Modeling of Large-Scale Behavioral Data. The successful use of context in various scientific and practical applications requires an integrative framework for understanding how diverse contextual factors and behavioral interventions shape choice. We have taken steps toward the development of such a framework by collecting a large experimental data set of choice behavior under the influence of many distinct contextual factors. Our data set is an order of magnitude larger than that used in most existing research and offers unparalleled insight into the computational mechanisms influenced by different behavioral interventions. We fit the parameters of the DDM (7, 8), a prominent model of choice process, on decisions under the influence of each contextual factor in our data set. The parameters of this model describe three core computational mechanisms underlying decision-making: predecisional bias (or equivalently baseline activation or statistical prior), drift rate (change in activation or use of evidence during choice deliberation) and decision boundary (activation threshold or acceptable error) (11).

Our results show that different interventions have selective influences on these three computational mechanisms. In doing so, they reveal nuanced insights about how the interventions alter choice probability and RT. For example, we found that providing social norms increased the choice probability for the normative option by altering both starting points and drift rates, whereas making an option a default increased choice probability by only altering starting points. These differences imply that social norms have an effect on preferences, but defaults do not. Prior work that measured intervention effects using only changes to choice probability cannot disentangle these selective influences. Such prior work also cannot measure the effects of interventions such as accountability, which increase the extent of deliberation but may not have a systematic effect on choice probability.

The power of our model relies on robust estimation of its various parameters. In the *SI Appendix*, we demonstrate that the parameters of our model can be accurately identified, and through cross-validation analysis, we show that the model makes good out-of-sample predictions of choice probabilities and RT distributions. We also show that replacing the DDM with an alternative evidence accumulation model did not alter our interpretation of the intervention effects. Our model is not too complex, but could it be too simple? Many additional cognitive mechanisms can be added to the simple DDM, such as between-trial variability in the parameters, collapsing boundaries, and nonlinear transformation of attribute values. We doubt that these extensions will change our main results, though they may reveal additional nuances in the effects of behavioral interventions on cognitive processes.

Model-Based Intervention Design. Our results offer academics and practitioners model-based methods for choosing between contextual influences based on the goals of the intervention. If the intervention aims to change people's preferences, then drift rate interventions like recommendations and social norms (as well as informational prompts), are suitable nudges. By contrast, defaults, option salience manipulations, and other interventions that have a selective effect on the starting point are better for altering choice probabilities without changing preferences. Starting point effects in the DDM diminish as decision boundaries increase, implying that the effects of starting point interventions are stronger for quick and error-prone decision makers and weaker for decision makers who are willing to deliberate extensively (4). Of course, some settings may require interventions that merely encourage extensive deliberation without systematically biasing choice in favor of either option. In such settings, our analysis would recommend decision boundary interventions, like accountability and rejection task framing.

The DDM model also offers a data-driven taxonomy of behavioral interventions in terms of their core computational mechanisms. The intervention space shown in Fig. 4 can, for example, quantify the similarities between different interventions, which, in turn, can be used to substitute one intervention for another or to develop novel interventions composed of two or more interventions. Consider, for example, combining an intervention that shifts the drift rate to favor a target option with an intervention that increases the decision boundary. The former would increase the choice probability of the target to above 50%. The latter would raise the threshold and thus reduce the probability of choosing the less preferred option. If we assume that intervention effects compose additively in the parameter space, the boundary intervention would amplify the effect of the drift rate intervention. More generally, choice probability depends on nuanced interactions among the three DDM parameters, and context effects that selectively influence different parameters can be combined to target the specific goals of the intervention. Although the effects of interventions on parameters may not be additive, and future experiments are necessary, our approach provides theory-driven quantitative predictions that can be used to motivate and design these experiments.

Many of the effects used to derive the recommendations in our framework have an empirical precedent. For example, our study replicates many prior findings regarding the effects of social cues, frames, defaults, attention, and emotion on decision-making. It also, however, provides a joint empirical investigation and integrative theoretical framework with which to compare and combine these contextual factors. Our framework is quantitative, and the insights derived from this framework come with precise predictions for choice probability and RT, allowing for rigor in the development and design of behavioral interventions.

Theory Integration across Disciplines. Our experiments involved a laboratory paradigm with multiple choices made by each individual. This methodology is popular in psychology and neuroscience, as it offers precise experimental manipulations, data on multiple decision variables like choice and RTs, as well as controls for individual heterogeneity in model fitting. Yet behavioral interventions typically target real-world choices, which may differ from those observed in laboratory experiments. For example, most everyday decisions are spread out over time, such that people make fewer pairwise choices within a day but encounter similar problems every so often. This could lead to different intervention effects, especially for starting point parameters, which are sensitive to experience. Therefore, field experiments—ideally mega-studies that jointly test a number of different behavioral interventions in a representative population of real-world decision makers (35–37)—should be conducted to test how our DDM-based conclusions generalize from laboratory experiments to field data.

It is also important to note that our approach adopts a fairly narrow conceptualization of the idea of a behavioral intervention and choice context. Behavioral interventions are typically understood to include not only features of the choice presentation (as with most of the interventions discussed in this article)

but also crucial components of the behavior system, such as people's opportunities and physical capabilities (see refs. 38 and 39 for a review and discussion). Additionally, behavioral interventions can influence behavior not only through nudges (4) but also through explicit incentives, training, persuasion, and even coercion. We suspect that many of these types of interventions will selectively influence the three DDM parameters. Testing this and, by doing so, extending our approach to model and interpret prominent behavioral change frameworks in implementation science is an important avenue for future work.

Conclusion

Our approach can be used to analyze a large number of policy-relevant behavioral interventions using a prominent cognitive and neurocomputational theory of choice. The parameters that we analyzed in this paper have direct interpretations in terms of neural variables and have been shown to respond to task and context manipulations across numerous domains in the cognitive and neural sciences, including perception, language, categorization, and memory. By relating prominent empirical regularities in the social and behavioral sciences to an established theoretical paradigm in the natural sciences, we offer a cohesive transdisciplinary approach to understanding human behavior.

Methods

Participants and Procedure. The study was approved by the ethics committee of the University of Pennsylvania. We recruited participants through various university experimental participant pools, with a combination of paid and unpaid participants to diversify our sample. All experiments took place on computers in a university behavioral laboratory, and the experiment lasted approximately 1 hour. Exp. 1 tested consumer choice and involved 608 participants (64.7% female, mean age 23.2 y, range 17–68 y) making choices between pairs of restaurants (each with a user rating and a price). Exp. 2 tested financial choice and involved 627 participants (69.6% female, mean age 21.9 y, range 18–65 y) making choices between pairs of health insurance plans (each with a deductible and a premium). Although we were constrained by our experimental paradigm to recruit participants who could take the experiment in the laboratory, our participant pools consisted of both college students and noncollege students, resulting in a relatively diverse sample of participants. All participants provided informed consent before the start of the experiment.

In exp. 1, we set the range of user ratings from 6 to 10 and the range of prices from 10 to 40. In exp. 2, we set the range of deductibles from 0 to 6,000 and the range of premiums from 30 to 380. We then randomly generated 80 unique choice problems within the predetermined range of attribute values. Additionally, we ensured that only a small set of problems included a dominating option (10/80 in exp. 1 and 5/80 in exp. 2; we excluded these problems when fitting the DDM). Before running the formal experiment in the laboratory, we tested the problem sets on an online experimental sample and found that choices in both experiments were fairly balanced.

Each experiment implemented 15 between-participant interventions. Participants in each experiment were randomly assigned to one of the between-participant interventions and made multiple choices both with and without the intervention. Thirteen between-participant interventions were implemented in both experiments. The other interventions involved money priming and food-quality priming in exp. 1 and informational prompts emphasizing the importance of either low deductibles or low premiums in exp. 2. This change was necessary as money and quality priming would not have different effects on the attributes in exp. 2 (both of which were monetary), and informational prompts would be redundant in exp. 1 (as participants have considerable prior experience with restaurant prices and user ratings). Finally, in addition to the 17 between-participant interventions across the two experiments, we also counterbalanced the presentation order of choice options and attributes in each experiment, allowing us to measure the effects of attribute and option ordering on choice within participants. This resulted in a total of 19 context-based interventions for our analysis.

The between-participant intervention implemented for each participant was randomly selected at the beginning of the session. The order of the baseline condition and the intervention condition was also randomly determined for each participant. To minimize the carry-over effect of the first-presented condition, we asked participants to complete a 5-min filler task between the two conditions. Thus, the experiment consisted of three total parts, with parts 1 and 3 corresponding to the baseline and intervention conditions (in random order) and part 2 corresponding to the filler task. The instructions for the baseline conditions for exp. 1 and exp. 2 can be found in *SI Appendix*, Table S1. The additional procedures and instructions for the intervention conditions can be found in *SI Appendix*, Table S2.

Each participant completed five practice trials in part 1 and part 3, 160 trials in the baseline condition (80 unique choice problems with option order counterbalanced), and 160 trials in the intervention condition (the same 80 unique choice problems as in the baseline condition with option order counterbalanced). The 160 trials in the 2 conditions were divided into 8 blocks of 20 trials each. There was a break after each block. Attribute order was held constant within a block and counterbalanced across blocks. All experimental procedures were preregistered with Open Science Framework (https://osf.io/y59jk/).

DDM. We grouped participants within each experiment based on which between-participant intervention condition they were assigned to, which resulted in 15 groups in each experiment. A hierarchical DDM was fit to each group using participants' choice and RTs recorded from the baseline condition, as well as the intervention condition. This hierarchical modeling approach estimates group- and individual-level parameters simultaneously, with group-level parameters (mean and SDs) forming the distributions from which individual participant estimates are sampled. We used a simple version of the DDM, with no between-trial variabilities in any of the parameters.

To analyze the two within-participant manipulations, we randomly selected 40 participants (who were assigned to any between-participant conditions) from both experiments and used their data recorded from the baseline condition to test the order effects. The attribute order effect was studied using the parameter difference between trials in which the quality attribute was presented above the price attribute (which we, for simplicity, refer to as the intervention condition) and trials in which the quality attribute was presented below the price attribute (the baseline condition). The option order effect was studied using the parameter difference between trials in which the high-quality option was presented to the left of the low-price option (the intervention condition) and trials in which the high-quality option was presented to the right of the low-price option (the baseline condition).

In our models, the upper boundary was associated with the choice of a high-quality option (i.e., the option with a higher score or a lower deductible in exp. 1 and exp. 2, respectively), and the lower boundary was associated with the choice of a low-price option (i.e., the option with a lower price or a lower premium in exp. 1 and exp. 2, respectively). In the baseline condition, the drift rate can be written as $v = \Delta U = U_1 - U_2$, where U_1 and U_2 denote subjective utility for the high-quality option and the low-price option, respectively. We followed the common practice of the field and modeled subjective utilities as weighted sums of the attribute values (40). Therefore, the utility difference between a pair of options is equal to the weighted sum of their attribute value differences. Without any context-based interventions, this is $v = \Delta U = v_{intercept} + v_{quality}\Delta Q + v_{price}\Delta P$. Here $v_{quality}$ and v_{price} denote multiplicative coefficients for the quality and price attributes, respectively. The larger their absolute values, the more important the attributes are to participants. The intercept, $v_{intercept}$, captures an overall stimuli-independent choice tendency. Positive $v_{intercept}$ captures a tendency to choose the high-quality option.

We captured the effects of interventions by allowing them to influence the starting point (z), the decision boundary (*a*), and the drift rate (*v*) of the DDM. For the starting point, we included a dummy variable to capture an additive shift between the intervention and the baseline condition. Moreover, for the attribute salience, option salience, default options, social norm, and experimenter recommendation conditions, we included another dummy variable to indicate which option was selectively (or asymmetrically) targeted by the manipulation (i.e., the option that dominated on the salient attribute, or the option that was salient,

default, popular, or recommended). We also allowed these two types of variables (additive shift between the intervention vs. baseline condition, and asymmetric effect of one of the options) to have an effect on the drift rate. In addition to these, we allowed the weights for the two attributes (quality and price) to be different under different conditions. Finally, we permitted an additive shift in the decision boundary between the two conditions. To control for the task order effect, we generated a dummy variable for each participant, indicating whether the condition was implemented in part 1 or part 3 of the experiment, and we allowed this variable to influence the decision boundary parameter (*a*). *SI Appendix*, Table S4 presents a summary of model specifications.

In summary, we assumed that the interventions had linear effects on all DDM components. We subsequently estimated these effects by fitting choice and RTs with the HDDMRegression function in HDDM (41), a Python package for hierarchical Bayesian estimation of DDMs. The Bayesian approach permits direct inferences for parameter posterior distributions. To fit the models, three chains of 20,000 samples were generated, where the first 5,000 were burn-ins, and a thinning of 2 was applied. Gelman-Rubin convergence statistics for model parameters were all close to 1, suggesting that the sample size was sufficient for the chains to converge.

Additional Analysis. In many cases, we used multiple parameters to capture the difference on a specific DDM component between the intervention condition and the baseline condition, and thus had to integrate their effects to produce Fig. 3. Here, we describe the procedure for generating the composite effect on the starting point first. For interventions without any asymmetric effects, this starting point effect is simply the group-level additive change of the starting point between the intervention condition and the baseline condition. For those interventions that selectively targeted (i.e., had an asymmetric influence) on one of the options (i.e., attribute salience, option salience, default options, social norm, experimenter recommendation conditions), the composite measure combines the group-level additive shift and asymmetric effect in all posterior samples. Because the asymmetric effect can either favor the high-quality option or the low-price option, two starting point effects were computed separately for these conditions. One of the starting point effects combined the additive shift and the asymmetric effect as if it was favoring the high-quality option, and the other combined the additive shift and the asymmetric effect as if it was favoring the low-price option (these effects are shown separately in Fig. 3).

To derive the drift rate effect of an intervention, we first took the mean differences of both the quality and price attributes across all the choice problems and used it to generate a typical choice problem. Based on the estimated group-level additive change, attribute weight change, and asymmetric shift between the intervention and the baseline conditions, we computed the overall drift rate change for this typical trial in all posterior samples, and we used their mean to measure the drift rate effect of an intervention. As the drift rate is a linear function of the attribute values, our procedure was equivalent to first computing the drift rates for all questions and then taking the mean of all those drift rates. As in the starting point, the asymmetric effect in the drift rate can favor either the high-quality option or the low-price option, and thus two drift rate effects were computed separately for the interventions with asymmetric effects. One of the drift rate effects combined the additive shift, the weight change, and the asymmetric effect favoring the high-quality option, and the other combined the additive shift, the weight change, and the asymmetric effect favoring the low-price option. Finally, the decision boundary effect was measured as the group-level additive shifts between intervention and baseline conditions.

To generate the three-dimensional space of behavioral interventions in Fig. 3*A*, we first scaled the DDM effects. This was achieved by dividing the posterior mean of the group-level composite effects by their associated SDs across the posterior samples. We then took these scaled effects' absolute values to reflect standardized intervention effect sizes on the DDM components. To account for interventions involving an asymmetric effect on the starting point or the drift rate, we averaged the standardized starting point or drift rate effect obtained from trials in which the highlighted option favored the high-quality or the low-price option. We used the Euclidean distance to calculate vector magnitudes and used Ward's minimum variance method to generate the hierarchical clustering in Fig. 4*C*.

Data Availability. Behavioral experiment data have been deposited in the Open Science Framework database (https://osf.io/y59jk/).

ACKNOWLEDGMENTS. Funding was received from NSF Grant SES-1847794.

1. A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).
2. S. DellaVigna, Psychology and economics: Evidence from the field. *J. Econ. Lit.* **47**, 315–372 (2009).
3. E. U. Weber, E. J. Johnson, Mindful judgment and decision making. *Annu. Rev. Psychol.* **60**, 53–85 (2009).
4. R. H. Thaler, C. R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin, 2009).
5. P. W. Glimcher, E. Fehr, Eds., *Neuroeconomics: Decision Making and the Brain* (Academic Press, 2013).
6. J. R. Busemeyer, S. Gluth, J. Rieskamp, B. M. Turner, Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends Cogn. Sci.* **23**, 251–263 (2019).
7. R. Ratcliff, P. L. Smith, A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.* **111**, 333–367 (2004).
8. J. I. Gold, M. N. Shadlen, The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
9. I. Krajbich, C. Armel, A. Rangel, Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).
10. R. Bogacz, E. Brown, J. Moehlis, P. Holmes, J. D. Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
11. R. Bogacz, Optimal decision-making theories: Linking neurobiology with behaviour. *Trends Cogn. Sci.* **11**, 118–125 (2007).
12. H. R. Heekeren, S. Marrett, L. G. Ungerleider, The neural systems that mediate human perceptual decision making. *Nat. Rev. Neurosci.* **9**, 467–479 (2008).
13. P. W. Schultz, J. M. Nolan, R. B. Cialdini, N. J. Goldstein, V. Griskevicius, The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* **18**, 429–434 (2007).
14. R. B. Cialdini, N. J. Goldstein, Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **55**, 591–621 (2004).
15. N. Harvey, I. Fischer, Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Hum. Decis. Process.* **70**, 117–133 (1997).
16. G. Bohner, N. Dickel, Attitudes and attitude change. *Annu. Rev. Psychol.* **62**, 391–417 (2011).
17. E. J. Johnson, D. Goldstein, Medicine. Do defaults save lives? *Science* **302**, 1338–1339 (2003).
18. B. C. Madrian, D. F. Shea, The power of suggestion: Inertia in 401(k) participation and savings behavior. *Q. J. Econ.* **116**, 1149–1187 (2001).
19. T. Mann, A. Ward, Attention, self-control, and health behaviors. *Curr. Dir. Psychol. Sci.* **16**, 280–283 (2007).
20. K. C. Armel, A. Beaumel, A. Rangel, Biasing simple choices by manipulating relative visual attention. *Judgm. Decis. Mak.* **3**, 396–403 (2008).
21. J. Baron, I. Ritov, Reference points and omission bias. *Organ. Behav. Hum. Decis. Process.* **59**, 475–498 (1994)
22. D. Kahneman, A perspective on judgment and choice: Mapping bounded rationality. *Am. Psychol.* **58**, 697–720 (2003).
23. E. Shafir, I. Simonson, A. Tversky, Reason-based choice. *Cognition* **49**, 11–36 (1993).
24. E. J. Johnson, A. Tversky, Affect, generalization, and the perception of risk. *J. Pers. Soc. Psychol.* **45**, 20 (1983).
25. N. Schwarz, G. L. Clore, Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J. Pers. Soc. Psychol.* **45**, 513 (1983).
26. M. L. Finucane, A. Alhakami, P. Slovic, S. M. Johnson, The affect heuristic in judgments of risks and benefits. *J. Behav. Decis. Making* **13**, 1–17 (2000).
27. A. L. Alter, D. M. Oppenheimer, Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev.* **13**, 219–235 (2009).
28. N. Mandel, E. J. Johnson, When web pages influence choice: Effects of visual primes on experts and novices. *J. Consum. Res.* **29**, 235–245 (2002).
29. C. Deck, S. Jahedi, The effect of cognitive load on economic decision making: A survey and new experiments. *Eur. Econ. Rev.* **78**, 97–119 (2015).
30. L. Guo, J. S. Trueblood, A. Diederich, Thinking fast increases framing effects in risky decision making. *Psychol. Sci.* **28**, 530–543 (2017).
31. J. W. Payne, J. W. Payne, J. R. Bettman, E. J. Johnson, *The Adaptive Decision Maker* (Cambridge University Press, 1993).
32. F. Lieder, T. L. Griffiths, Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, e1 (2020).
33. J. S. Lerner, P. E. Tetlock, Accounting for the effects of accountability. *Psychol. Bull.* **125**, 255–275 (1999).
34. M. H. Birnbaum, J. W. Jou, A theory of comparative response times and "difference" judgments. *Cognit. Psychol.* **22**, 184–210 (1990).
35. K. L. Milkman *et al.*, Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
36. K. Muralidharan, P. Niehaus, Experimentation at scale. *J. Econ. Perspect.* **31**, 103–124 (2017).
37. S. DellaVigna, E. Linos, RCTs to scale: Comprehensive evidence from two nudge units. *NBER Working Paper* (2020). http://doi.org/10.3386/w27594.
38. S. Michie, M. M. van Stralen, R. West, The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implement. Sci.* **6**, 42 (2011).
39. P. Nilsen, "Making sense of implementation theories, models, and frameworks," in *Implementation Science 3.0*, B. Albers, A. Shlonsky, R. Mildon, Eds. (Springer, Cham, 2020), pp. 53–79.
40. R. L. Keeney, H. Raiffa, *Decisions With Multiple Objectives: Preferences and Value Trade-Offs* (Cambridge University Press, 1993).
41. T. V. Wiecki, I. Sofer, M. J. Frank, HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Front. Neuroinform.* **7**, 14 (2013).