PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Model-based algorithms shape automatic evaluative processing

David E. Melnikoff[a,1] and Benedek Kurdi[b,1]

Edited by Elke Weber, Princeton University, Princeton, NJ; received August 21, 2024; accepted May 1, 2025

Computational theories of reinforcement learning suggest that two families of algorithm—model-based and model-free—tightly map onto the classic distinction between automatic and deliberate systems of control: Deliberate evaluative responses are thought to reflect model-based algorithms, which are accurate but computationally expensive, whereas automatic evaluative responses are thought to reflect model-free algorithms, which are error-prone but computationally cheap. This framework has animated research on psychological phenomena ranging from habit formation to social learning, moral decision-making, and cognitive development. Here, we propose that model-based and model-free algorithms may not be as aligned with deliberate and automatic evaluative processing as prevailing theories suggest. Across three preregistered behavioral experiments involving adult human participants (total n = 2,572), we show that model-based algorithms shape not only deliberate but also automatic evaluations. Experiment 1 numerically replicates past findings suggesting that deliberate (but not automatic) evaluative responses are uniquely shaped by model-based algorithms but, critically, also reveals confounds that render interpretation of this evidence equivocal. Experiments 2 to 3 eliminate these confounds and reveal robust model-based contributions to automatic evaluative processing across two measures of automatic evaluation, supported by multinomial processing tree modeling. Together, these results suggest that dominant frameworks may considerably underestimate both the ubiquity of model-based algorithms and the computational sophistication of automatic evaluative processing.

reinforcement learning | model-based control | model-free control | evaluation | automaticity

Significance

One of the oldest and most influential views in psychology is that automatic and deliberate responses rely on qualitatively distinct algorithms. The most recent incarnation of this idea comes from computational reinforcement learning, which suggests that deliberate evaluative processing reflects accurate but computationally intensive "model-based" algorithms, whereas automatic evaluative processing reflects error-prone but computationally efficient "model-free" algorithms. We challenge this framework by showing that model-based algorithms shape both deliberate and automatic evaluations. This finding reveals that model-based algorithms are more pervasive, and automatic processes more computationally sophisticated, than previously thought. It also casts doubt on the dominant idea that automatic processes are inherently more error-prone than deliberate ones.

According to computational theories of reinforcement learning (1–4), decision-makers rely on at least two types of algorithm to evaluate and choose between competing options. There are model-based algorithms, which use a representation of the environment's causal structure to determine which option best serves the decision-maker's current goals. Then there are model-free algorithms, which evaluate options based solely on how rewarding they have been in the past.

Model-free algorithms are computationally cheaper, but more rigid and error-prone. They primarily change slowly through error-based learning—the process of repeatedly choosing an option, experiencing unexpected outcomes, and incrementally updating based on the resulting error signals. While model-free evaluations can change quickly, this ability is highly circumscribed, occurring only when an option is reassigned to a new conceptual category with a different reward history (5, 6). For instance, if you learned that a wine you thought was Chardonnay was actually Riesling, your model-free evaluation of the wine would immediately shift to reflect your history with Rieslings. Beyond this, however, model-free algorithms lack any mechanism for rapid updating.

Model-based algorithms are far more flexible. They swiftly adapt to changing goals and environmental contingencies, often making them more accurate than their model-free counterparts. However, according to prevailing theories, model-based algorithms contribute only to a subset of the evaluations that guide people's decisions: those that are made deliberately rather than automatically (2, 7–13). By contrast, automatic evaluations are considered model-free.

To illustrate, imagine that when driving to refuel your car, you usually choose between two highway exits: Exit 1 and Exit 2. You prefer Exit 1 because it gets you to Gas Station 1, which almost always has the lowest prices. However, today you learned that the lowest prices are at Gas Station 2, which is located off Exit 2. Upon receiving this information, your model-based evaluations of the exits would immediately change in favor of Exit 2, whereas your model-free evaluations would remain the same. So, when deciding which exit to take, model-based algorithms would serve you best. However, leading theories suggest that to use model-based computations, you would need to evaluate the exits deliberately; if you relied on your automatic evaluations—those that spring to mind spontaneously—you would fall back on model-free computations and, in this scenario, take the suboptimal exit.

Author affiliations: [a]Graduate School of Business, Stanford University, Stanford, CA 94305; and [b]Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL 61820

Author contributions: D.E.M. and B.K. designed research; performed research; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: dmelnik@stanford.edu or kurdi@illinois.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2417068122/-/DCSupplemental.

Published June 20, 2025.

This standard portrayal of evaluative processing may need revising. We propose that model-based algorithms shape not only deliberate evaluations but automatic evaluations as well. If this proposal is correct, then model-based algorithms are far more ubiquitous, and automatic evaluations considerably more computationally sophisticated, than previously thought. Entrenched accounts of when and why decision-makers fail to use model-based algorithms would come under question, as would the influential view that model-based and model-free algorithms implement two distinct and competing systems of control, one automatic and the other deliberate (2, 7, 12, 13).

Though our hypothesis contradicts dominant theorizing, it is not entirely without support. For instance, model-based processing has been found to reduce self-reported reliance on mental effort (14), which could, in theory, result from model-based algorithms operating automatically. Additionally, according to some perspectives, model-based and model-free algorithms do not compete for control over evaluative processing, as traditionally assumed, but instead collaborate in ways that could, in principle, allow model-based algorithms to shape automatic evaluations (15–19). Still, the theoretical record overwhelmingly suggests that model-based algorithms shape only deliberate evaluative processing, and every direct test of this perspective has been interpreted as supporting it (11–13, 20).

The rest of this paper proceeds as follows. First, we critically examine the evidence against our claim that model-based algorithms support automatic evaluation. Our analysis will reveal, perhaps surprisingly, that only one set of findings seriously challenges our hypothesis. We meet this challenge by offering an alternative account of these findings, which we empirically verify. Finally, we directly support our hypothesis with experimental evidence that model-based algorithms indeed shape automatic evaluations.

Evidence Against Model-Based Input to Automatic Evaluation

Here, we present exhibits A and B in the case against model-based input to automatic evaluative processing. Exhibit A comprises work showing that model-based algorithms selectively fail when cognitive resources are constrained (12, 13, 16, 20). Factors like working memory load and stress have been found to reduce or even eliminate model-based input to decision-making while leaving model-free input intact. Such findings have been taken to support two conclusions. One is that model-based algorithms demand more cognitive resources than model-free algorithms. This is not in question. The other conclusion is that model-based algorithms shape deliberate but not automatic evaluative processing. Of this, we are skeptical. The underlying assumption here is that cognitive constraints impair deliberate processes but not automatic processes. This assumption is not quite right: Cognitive constraints can selectively impair automatic processing (21–23).

One of many examples comes from the Stroop task (24), the archetypical assay of automatic versus deliberate processing. In the Stroop task, the automatic process of word-reading interferes with the deliberate process of color-naming. However, working memory load can selectively impair the automatic word-reading process, eliminating the Stroop effect completely (25). The upshot: Deleterious effects of working memory load on model-based processing do not preclude model-based input to automatic evaluation. Whether a process is automatic, and whether it fails under working memory load, are separate issues (21–23). We can dismiss exhibit A.

Exhibit B poses a greater challenge. It comprises evidence from a series of studies (11) that combined classic reinforcement learning tasks with the Implicit Association Test (IAT) (26)—a standard measure of automatic evaluations. To show how these studies might be reconciled with our hypothesis, we must describe them in some detail.

To start, participants repeatedly chose between members of two fictitious social groups: Group A and Group B. Whenever participants chose Group A, they saw a shape—Shape A—and subsequently received points. Whenever participants chose Group B, they saw a different shape—Shape B—and subsequently lost points. Next, some participants learned that the transition structure had changed: Shape A now transitioned to losses, and Shape B to gains. This change meant that Group B would now yield gains, and Group A would now yield losses. The new contingencies between choice options and point outcomes were not experienced directly—they had to be inferred from the updated transition structure.

This type of manipulation, where a distal change in transition structure (e.g., from shapes to points) alters the value of proximal choice options (e.g., social groups), is called transition revaluation (Fig. 1). It is designed to dissociate model-free from model-based processing. Model-based algorithms can immediately revaluate proximal choice options based on distal changes in transition structure, but model-free algorithms cannot. So, if model-based computations do not shape automatic evaluations, then automatic evaluations should not respond to transition revaluation.

This is what the experimenters observed (11). To confirm successful learning of the new transition structure, the experimenters instructed participants to choose the social group that would yield points. Most participants chose correctly, revealing successful learning. However, this learning did not affect automatic evaluations of the social groups as assessed by the IAT. IAT scores revealed an automatic preference for Group A, and the strength of this preference was unaltered by transition revaluation. A straightforward interpretation of these findings—and the one favored by the original authors—is that automatic evaluations receive no model-based input.

Here, we offer a different interpretation. Recent work shows that transition revelation procedures can underestimate model-based processing, especially when participants interpret the procedure differently than intended (14, 27). Drawing on this work, we propose that participants interpreted the transitions from social groups to shapes as evidence of the groups' moral character. On our account, participants believed that Group A deliberately generated Shape A to confer gains, and that Group B deliberately generated Shape B to confer loses, which led participants to deem Group A morally superior to Group B (28–35). This inference, we suggest, persisted after transition revaluation because the transition structure was altered when neither group was present—behind their backs, so to speak—making the change appear nondiagnostic of the groups' intentions (28).

If participants indeed considered Group A morally superior both before and after transition revaluation, the failure of transition revaluation to alter automatic evaluations is just as consistent with model-based processing as it is with model-free processing. Any plausible model of the social world would expect moral groups to yield far more value in the long run than immoral groups, producing a model-based preference for the former (36–38). This preference should be strong—strong enough to remain essentially unaltered by a manipulation that makes a moral group predictive of point gains and an immoral group predictive of point losses. Compared to the long-run value of moral character over immoral character, the value of meaningless points is negligible.

So, why did transition revaluation fail to alter automatic evaluations? Perhaps because automatic evaluations were based on a

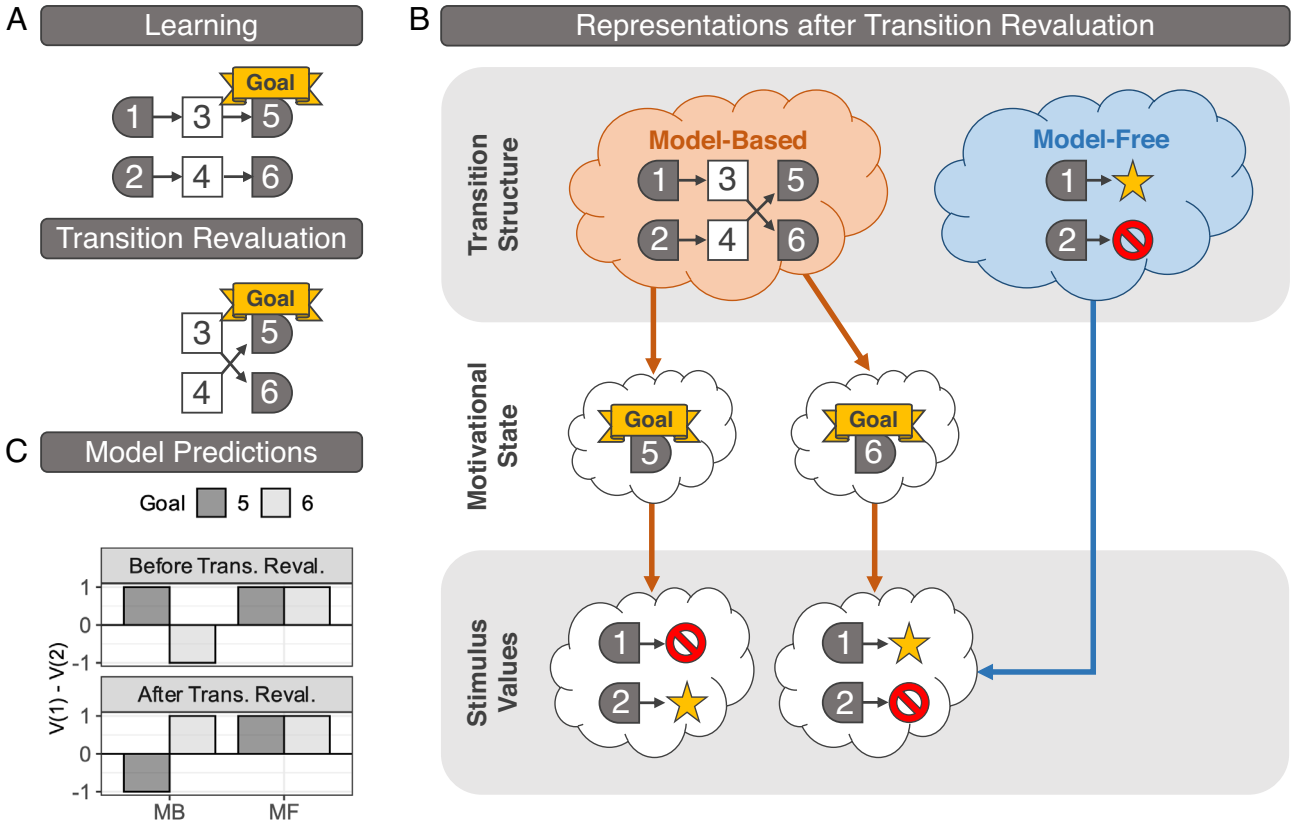

Fig. 1. Schematics for learning trials, mental representations, and model predictions in transition revaluation procedures. (*A*) Schematics for initial learning trials (*Top* panel) and transition revaluation manipulation (*Bottom* panel) in a transition revaluation procedure. Each number denotes a unique state, and each arrow denotes a deterministic transition. The "goal banner" denotes the desired terminal state. (*B*) Schematics for mental representations of the transition structure (*Top* panel) and stimulus values (*Bottom* panel) as a function of motivational state (*Middle* panel). Positive values are denoted by a star and negative value by a "no" symbol. A model-based agent (red) uses a representation of the full transition structure to generate a preference for whichever initial state leads to its current goal. A model-free agent (blue) evaluates the initial states based on cached values from the initial learning phase, which could not be updated during the transition revaluation procedure. Accordingly, the model-free agent's evaluations are insensitive to the transition revaluation procedure as well as its motivational state. (*C*) Model predictions for the model-based (MB) and model-free (MF) agent as a function of motivational state. V(1) denotes the value of State 1 and V(2) denotes the value of State 2; V(1)–V(2) denotes the preference for State 1 over State 2.

causal model that favors moral groups over immoral groups, and Group A was deemed morally superior both before and after transition revaluation. Put another way, transition revaluation may have failed to alter automatic evaluations not because automatic evaluations lack model-based input, but because the model-based input was essentially the same across conditions.

The Present Project

We conducted three tests of the hypothesis that model-based algorithms shape automatic evaluations. Experiment 1 tested our alternative account of existing evidence that automatic evaluations are insensitive to transition revaluation. The results support our speculation that, in past work (11), transition revaluation failed to alter automatic evaluations not because automatic evaluations are model-free, but because participants drew unintended inferences about the experimental stimuli. In Experiments 2 to 3, we introduced a new transition revaluation paradigm—one modified to prevent the sort of inferences that thwarted the original. Across multiple measures of automatic evaluation, these experiments provide strong evidence that automatic evaluations reflect model-based processing.

Experiment 1

Experiment 1 replicated the transition revaluation paradigm that failed to alter automatic evaluations in previous work (11). The only change was the addition of two dependent measures. One was a measure of moral judgments. We hypothesized that the group that initially predicted gains would be considered morally superior to the group that initially predicted losses, and that the strength of this effect would be unmoderated by transition revaluation.

The second addition was a self-report measure of deliberate evaluations. Since deliberate evaluations are largely model-based, we used them to test our assumption that model-based preferences in this paradigm reflect moral judgments. If our assumption is correct, deliberate evaluations should follow the same pattern as moral judgments.

The paradigm proceeded in two stages (for full details, see *SI Appendix*). In stage 1, participants learned the transition structure between two fictional social groups (Group A and Group B), two intervening stimuli (Shape A and Shape B), and two outcomes (+5 Points and −5 Points). Depending on condition, Group A was called either "Niffians" or "Laapians" (the alternative of the two names was used for Group B); Shape A was either a horizontal line or a vertical line (the alternative of the two shapes was used for Shape B). On each of 20 learning trials, participants chose one of the two social groups and then observed two transitions, the first to an intervening stimulus and the second to an outcome. The transitions were deterministic. Group A always led to Shape A, which always led to +5 Points; Group B always led to Shape B, which always led to −5 Points. Participants were instructed to earn as many points as possible. They made their choices using the E and I keys of their keyboard; which key corresponded to which social group was randomized on each trial. Stage 1 concluded with

four additional binary choices between the social groups, again aimed at earning points. Of these four choices, we used the total made in favor of Group A to assess whether participants learned the transition structure; those who had should choose Group A more often than not.

Stage 2 came next. What happened here depended on the condition to which participants were assigned: baseline learning or revaluation. In the baseline learning condition, participants immediately completed three dependent measures in the following order: a) an Implicit Association Test (IAT) to probe automatic evaluations of the social groups, b) another measure of deliberate evaluations, and, finally, c) another measure of moral judgments.

In the revaluation condition, participants passively observed a new transition structure between intermediate stimuli and outcomes (the social groups were not observed). In a complete reversal of the original contingency, Shape A now led to −5 Points and Shape B to +5 Points. After 20 trials of passive learning, participants made another four choices between the social groups to assess learning of the new transition structure. Then participants completed the same dependent measures from the baseline learning condition (IAT, deliberate evaluations, and moral judgments, in that order).

At this point, readers might wonder why we included a separate measure of deliberate evaluations. Why not infer deliberate evaluations from the choices participants made between the social groups? After all, choices often reflect preferences, and these choices were made deliberately; they were even termed "explicit evaluations" by the researchers who developed this paradigm (11). Our answer is that deliberate evaluations cannot be inferred from participants' choices because participants were instructed to choose the group that would lead to points—not the group they preferred. To comply with their instructions, participants may have chosen one group despite preferring the other. In fact, this is precisely what we expected to occur in the revaluation condition. Participants in this condition should, according to our hypothesis, consider Group A morally superior to Group B, and therefore preferable, but select Group B to fulfill the experimenter's request. Accordingly, deliberate evaluations were measured separately from participants' choices, which were used to confirm successful learning of the transition structure.

Participants Successfully Learned the Transition Structure. Participants successfully learned the initial contingencies between groups and outcomes (Fig. 2): They systematically chose Group A at the end of stage 1, $t(750) = 24.03$, $P < 0.0001$, Cohen's $d = 0.88$, $BF_{10} = 1.93 \times 10^{91}$. When the contingencies changed, participants successfully updated their beliefs. After transition revaluation, participants chose Group A less than they did before transition revaluation, $t(556.60) = 8.44$, $P < 0.0001$, Cohen's $d = 0.43$, $BF_{10} = 1.01 \times 10^{12}$, and less than participants in the baseline learning condition, $t(658.02) = 7.78$, $P < 0.0001$, Cohen's $d = 0.58$, $BF_{10} = 4.37 \times 10^{11}$. Not unexpectedly, this effect was a partial one. After transition revaluation, participants still preferentially selected Group A, $t(359) = 2.89$, $P = 0.004$, Cohen's $d = 0.15$, $BF_{10} = 3.58$, but at a reduced rate. We return to this finding later. For now, what matters is that the transition revaluation manipulation served its purpose of shifting beliefs about the contingencies between social groups and outcomes. Did automatic evaluations change as a result?

Automatic Evaluations Were Insensitive to Transition Revaluation. Replicating past work, transition revaluation had no effect on automatic evaluations, as indexed by the IAT (Fig. 2), $t(741.99) = -0.49$, $P = 0.621$, Cohen's $d = -0.04$, $BF_{01} = 10.88$, which always favored Group A, $t(750) = 6.33$, $P < 0.0001$, Cohen's $d = 0.23$, $BF_{10} = 1.12 \times 10^{7}$. This might seem to suggest that model-based algorithms have little to no influence on automatic evaluations. However, further analyses cast doubt on this interpretation.

Model-Based Input to Automatic Evaluation Cannot be Ruled Out. As we suspected, participants deemed Group A morally superior to Group B, $t(734) = 2.57$, $P = 0.01$, Cohen's $d = 0.09$, $BF_{10} = 1.09$ (Fig. 2). The strength of this attribution was unrelated to condition, $t(703.79) = 0.23$, $P = 0.818$, Cohen's $d = 0.02$, $BF_{01} = 11.81$. According to our hypothesis, the belief in Group A's moral superiority produced a model-based preference for Group A that was insensitive to transition revaluation.

Supporting this idea, morality judgments had a significant, positive effect on deliberate evaluations (which are largely

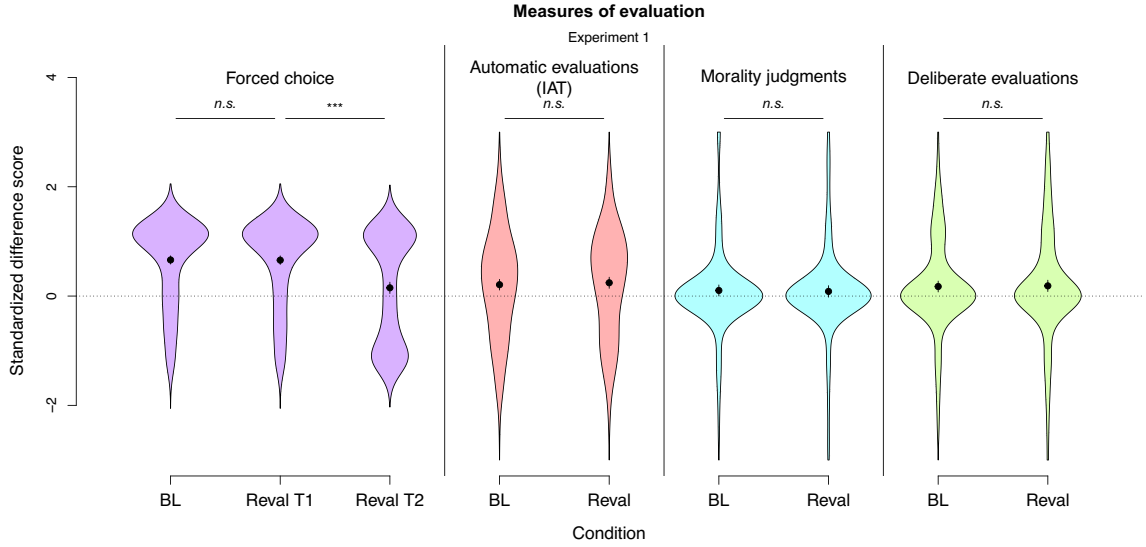

Fig. 2. Experiment 1 ($n = 751$): The distribution of forced choices, automatic evaluations (IAT scores), morality judgments, and deliberate evaluations by learning condition. The x-axis shows condition (control vs. revaluation), and the y-axis shows standardized difference scores between Group A (the group that initially predicted gains) versus Group B (the group that initially predicted losses); positive scores indicate a preference for Group A over Group B. The solid dots denote condition means, and the error bars show 95% CIs. All dependent measures were collected once (at the end of the final learning phase) except for choices in the revaluation condition, which were measured twice: at the end of the final learning phase (T2) and at the end of the initial learning phase (T1). BL = baseline learning; Reval = revaluation; *n.s.* = not significant; ***$P < 0.001$.

model-based), $\beta = 0.62$, $t(730) = 21.46$, $P < 0.0001$. Deliberate evaluations also followed the same qualitative pattern as morality judgments: They favored Group A, $t(736) = 4.98$, $P < 0.0001$, Cohen's $d = 0.09$, $BF_{10} = 1.09$, and the strength of this preference did not differ by condition, $t(675.75) = -0.16$, $P = 0.87$, Cohen's $d = 0.01$, $BF_{10} = 11.99$.

Could this consistent, model-based preference for Group A explain why participants automatically preferred Group A across conditions? Exploratory analyses point in this direction. IAT scores were significantly predicted by deliberate evaluations, $\beta = 0.36$, $t(735) = 10.32$, $P < 0.0001$, by morality judgments, $\beta = 0.27$, $t(733) = 7.73$, $P < 0.0001$, and by time-1 forced choices, $\beta = 0.16$, $t(749) = 4.41$, $P < 0.0001$, but not by time-2 forced choices, $\beta = 0.02$, $t(358) = 1.23$, $P = 0.220$.

In addition to preserving the possibility of model-based input to automatic evaluation, these findings help explain why transition revaluation reduced, but did not reverse, participants' tendency to select Group A. Despite being instructed to choose the group that would lead to points, participants may have based their choices, to some extent, on their deliberate preferences and moral judgments. Exploratory analyses support this conjecture. The tendency to choose Group A was an increasing function of deliberate preferences for Group A, $\beta = 0.09$, $t(735) = 2.33$, $P = 0.02$, and the belief in Group A's moral superiority at the time of choice, $\beta = 0.10$, $t(735) = 2.85$, $P = 0.005$.

To summarize, the current paradigm cannot rule out model-based input to automatic evaluation. Morality judgments consistently favored Group A, and were associated with a model-based preference for Group A that was insensitive to transition revaluation. Accordingly, the failure of transition revaluation to shift automatic evaluations need not imply that automatic evaluations lack model-based input; it is just as consistent with model-based processing as it is with model-free processing. The possibility that model-based algorithms shape automatic evaluations is back in play.

Experiment 2

If our hypothesis is correct, then transition revaluation should shift automatic evaluations, provided that participants refrain from drawing the sorts of character inferences documented in Experiment 1. We tested this prediction in Experiment 2. Specifically, we modified the original procedure by replacing the social groups with nonsocial stimuli to discourage attributions of mentalistic content (Fig. 3; for full details, see *SI Appendix*).

Experiment 2 also included an additional assay of model-based processing. Model-based algorithms can revaluate proximal stimuli not only in response to distal changes in transition structure, but also in response to changes in motivation (Fig. 1). Therefore, in addition to a transition revaluation manipulation, we included a motivation revaluation condition, in which we altered the outcome that participants aimed to obtain. Both manipulations should shift automatic evaluations if automatic evaluations receive model-based input.

In principle, automatic evaluations could shift in response to one manipulation but not the other. Such a result would imply that automatic evaluations are neither model-based nor model-free, but instead emerge from an intermediate class of algorithm (18, 39). For instance, algorithms based on the so-called successor representation (SR) respond to motivation revaluation (making them more flexible than model-free algorithms) but not to transition revaluation (making them less flexible than model-based algorithms). However, we find no evidence of SR-based algorithms or any other intermediate algorithm. Therefore, our discussion of this distinction ends here.

The experiment proceeded as follows. To start, all participants learned that they would play a game called "Oink Fest," the goal of which was to win first prize in a pig growing contest. To win, participants needed to help their "scrawny piglet" grow as large as possible.

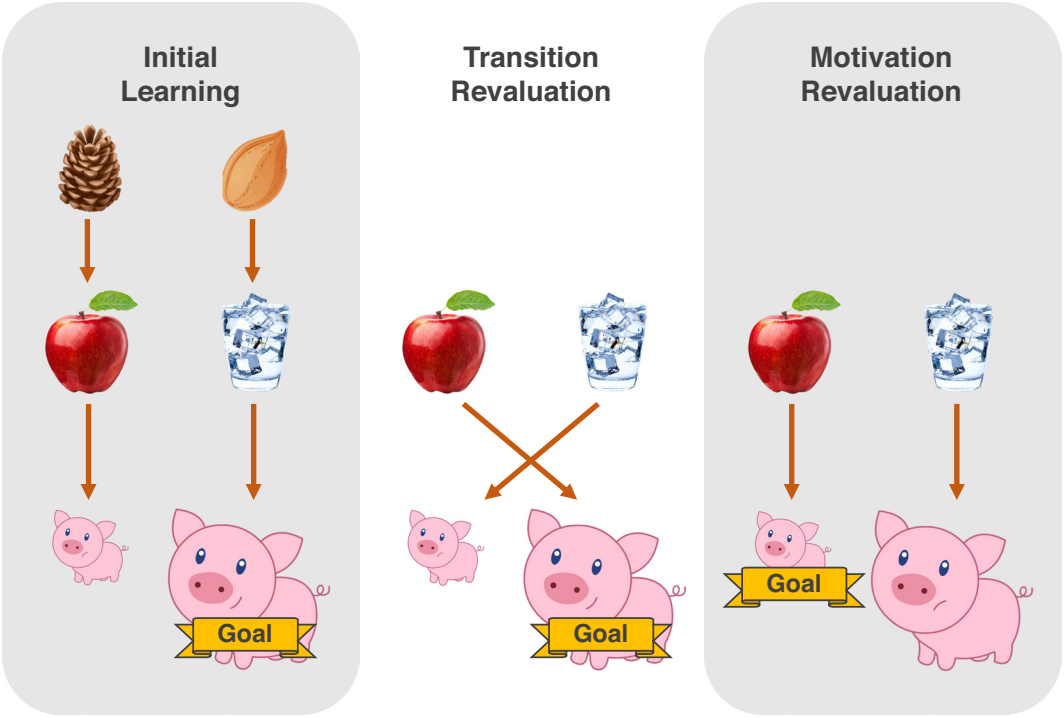

Fig. 3. Schematics of the procedure of Experiment 2. In the initial learning phase (*Left* panel), stage 1 stimuli (seeds and pinecones) transitioned deterministically to Stage 2 stimuli (food and water), which transitioned deterministically to terminal outcomes (smaller pig or bigger pig). Pairings between State 1 stimuli and Stage 2 stimuli were randomized between subjects, as were pairings between Stage 2 stimuli and terminal outcomes. The goal in the learning phase was always to grow the pig as big as possible. In the transition revaluation condition (*Middle* panel), the transitions between Stage 2 stimuli and terminal outcomes were reversed. In the motivation revaluation condition (*Right* panel), the goal was changed.

The piglet was said to need one of two resources to grow: food (because it was hungry) or water (because it was thirsty). Participants learned that the piglet would grow (Outcome A) each time it received its required resource (Resource A), and that it would shrink (Outcome B) each time it received the other resource (Resource B). To deliver resources to the piglet, participants needed to activate a machine, which ran on two types of "energy sources": pinecones and seeds. Participants learned that one of the energy sources (Energy Source A) would cause the machine to dispense Resource A, and the other energy source (Energy Source B) would cause the machine to dispense Resource B. To summarize, Energy Source A (pinecones or seeds, depending on condition) transitioned to Resource A (food or water, depending on condition), which transitioned to Outcome A (grow); Energy Source B transitioned to Resource B, which transitioned to Outcome B (shrink).

Participants completed 20 trials of Oink Fest. On each trial, they chose between Energy Sources A and B and observed the consequences of their choice: the machine dispensing Resource A or B followed by their piglet growing or shrinking. Participants made their choices using the E and I keys of their keyboard; which key corresponded to which energy source was randomized on each trial. After 20 trials, participants were informed that their pig got second prize. It was then announced that participants would complete a second round of Oink Fest.

In the baseline learning condition, participants learned that the second round was identical to the first. In the motivation revaluation condition, participants learned that in the second round, they would pursue a different goal: to make their pig as small as possible, rather than as big as possible. Specifically, these participants learned that in the second round, first prize would go to the "tiniest piglet," making Energy Source B the most valuable energy source (because it would lead to Resource B, which would make the piglet shrink). Finally, in the transition revaluation condition, participants learned that the contingencies between resources (food and water) and outcomes (growing vs. shrinking) had reversed: Resource A would now lead to Outcome B and Resource B to Outcome A. This manipulation, like the motivation revaluation manipulation, reversed which energy source was most valuable from Energy Source A to Energy Source B.

In the two revaluation conditions, participants then passively observed 10 transitions from resources to outcomes. Specifically, participants in the transition revaluation condition observed the new contingencies; participants in the motivation revaluation condition observed the same contingences from the first round, along with a message on each trial that communicated whether the specific outcome was desirable or not. During this phase, participants neither saw nor chose between energy sources.

Participants did not actually complete the second round of Oink Fest in any condition. Instead, they proceeded directly to the dependent measures, which included the same measures of automatic and deliberate evaluations of the proximal stimuli (pinecones vs. seeds) used in Experiment 1. The measure of morality judgments was omitted from this experiment (we assumed that pinecones and seeds were not regarded as moral agents), as was the behavioral measure of successful learning (we instead used self-report items to confirm that participants understood the transition structure and goal of each round).

Model-Based Preferences Were Successfully Manipulated. Both revaluation conditions succeeded at shifting model-based preferences as indexed by deliberate evaluations (Fig. 4). Specifically, deliberate evaluations varied by condition, $F(2, 962) = 37.17$, $P < 0.0001$, $\eta^2 = 0.07$, $BF_{10} = 1.84 \times 10^{13}$, such that they favored Energy Source A more in the baseline learning condition ($M = 24.56$, $SD = 35.65$) than in the motivation revaluation condition ($M = 8.88$, $SD = 24.58$), $t(962) = 6.58$, $P < 0.0001$, and the transition revaluation condition ($M = 5.77$, $SD = 28.49$), $t(962) = 8.01$, $P < 0.0001$. Deliberate evaluations did not differ across the two revaluation conditions, $t(962) = 1.28$, $P = 0.201$.

In the revaluation conditions, deliberate preferences for Energy Source A were reduced, but not reversed. Across all three conditions, participants maintained a significant deliberate preference for Energy Source A, $t(964) = 13.39$, $P < 0.0001$, Cohen's $d = 0.43$, $BF_{10} = 1.39 \times 10^{34}$. This result could imply that deliberate evaluations were mixtures of model-based computations (making them sensitive to revaluation) and model-free computations (preventing their complete reversal); alternatively, it is possible that model-based preferences themselves did not completely reverse. Either way, Experiment 2 succeeded where Experiment 1 did not: model-based preferences were manipulated, allowing for a critical test of whether model-based algorithms shape automatic evaluations.

Model-Based Algorithms Shaped Automatic Evaluations. Automatic evaluations, as measured by the IAT, followed the same pattern as deliberate evaluations (Fig. 4). Specifically, IAT scores differed across conditions, $F(2, 986) = 16.58$, $P < 0.0001$, $\eta^2 = 0.03$, $BF_{10} = 9.27 \times 10^4$, such that they favored Energy Source A more in the baseline learning condition ($M = 0.38$, $SD = 0.43$) than in the motivation revaluation condition ($M = 0.22$, $SD = 0.46$), $t(986) = 4.78$, $P < 0.0001$, or in the transition revaluation condition ($M = 0.21$, $SD = 0.46$), $t(986) = 5.07$, $P < 0.0001$. The two revaluation conditions did not differ from each other, $t(986) = 0.22$, $P = 0.825$. In absolute terms, IAT scores favored Energy Source A across all three conditions, $t(988) = 19.30$, $P < 0.0001$, Cohen's $d = 0.61$, $BF_{10} = 1.01 \times 10^{67}$. This preference was merely reduced in the two revaluation conditions, as was the deliberate preference. All results remained the same when analyses were repeated only using data from participants who passed all manipulation checks.

At a first approximation, these data suggest that automatic evaluations are sensitive to model-based processing. However, this conclusion hinges on the assumption that the IAT is a measure of automatic, rather than deliberate, evaluations. This assumption has been challenged by evidence suggesting that IAT responding emerges from a combination of relatively automatic and relatively controlled processes. As such, in exploratory analyses reported in full detail in *SI Appendix*, we used Bayesian hierarchical modeling (40, 41) to fit the quadruple process model (42)—a form of multinomial processing tree modeling (43)—to the IAT data from Experiment 2.

Quad modeling uses patterns of errors across trial types on the IAT to decompose responding into four processes: a) the automatic activation of positive and negative information, the controlled processes of b) detecting the correct response and c) overcoming any conflict between the correct response and automatic evaluation, as well as d) guessing. If the revaluation manipulations exert their effects via automatic evaluative processing, a)—but not b) or c)—should differ across the baseline learning and revaluation conditions.

This is exactly the pattern that we obtained. Compared to the baseline learning condition, Energy Source A automatically activated significantly less positive information in both revaluation conditions [motivation revaluation: $b = -0.0123$, 95% HDI (−0.0230; −0.0047); transition revaluation: $b = -0.0116$, 95% HDI (−0.0220; −0.0042)]. The two revaluation conditions did not differ from each other, $b = -0.0008$, 95% HDI (−0.0055; 0.0036). Similarly, Energy Source B automatically activated significantly less negative information in the motivation revaluation condition compared to the baseline learning condition, $b = -0.0052$, 95% HDI (−0.0122; −0.0007). A nonsignificant trend in the same direction emerged in the transition revaluation

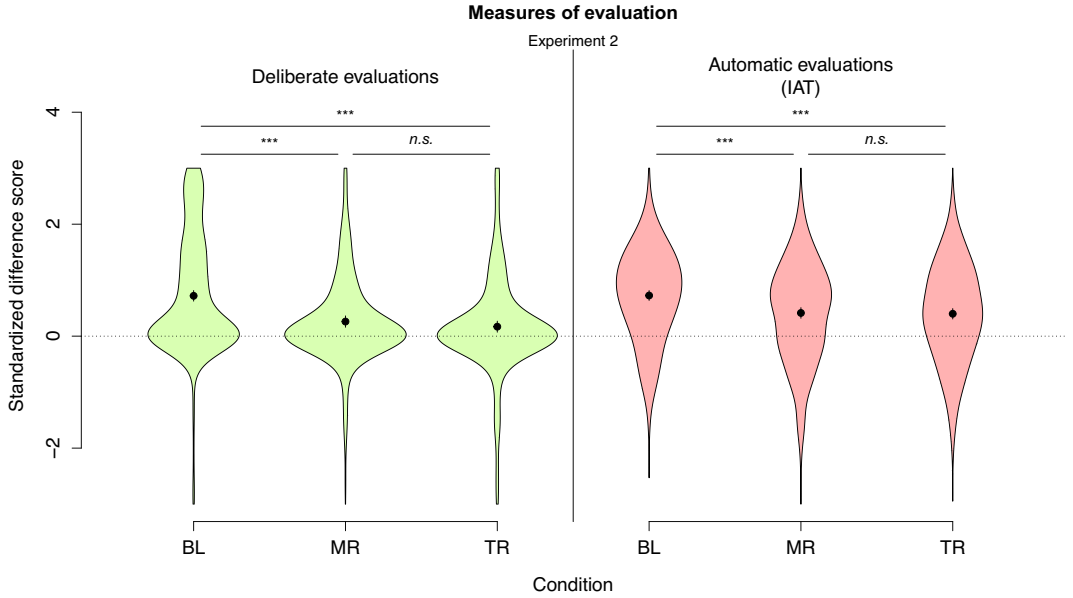

Fig. 4. Experiment 2 (n = 989): The distribution of deliberate evaluations and automatic evaluations (IAT scores) by learning condition. The x-axis shows condition (baseline learning, motivation revaluation, and transition revaluation), and the y-axis shows standardized difference scores between Energy Source A (which was initially rewarding) and Energy Source B (which was initially nonrewarding); positive scores indicate a preference for Energy Source A over Energy Source B. The solid dots denote condition means, and the error bars show 95% CIs. BL = baseline learning; MR = motivation revaluation; TR = transition revaluation; *n.s.* = not significant; ***P < 0.001.

condition, b = −0.0035, 95% HDI (−0.0098; 0.0010), with no difference across the two revaluation conditions, b = −0.0017, 95% HDI (−0.0062; 0.0020).

Neither revaluation manipulation had a significant effect on any of the remaining (controlled) parameters. This suggests that both revaluation conditions altered IAT performance exclusively by changing automatic evaluations, reinforcing the conclusion that automatic evaluations can reflect model-based processing.

Experiment 3

Both traditional and multinomial processing tree analyses of the data from Experiment 2 suggest that model-based algorithms can influence automatic evaluations as measured by the IAT. However, the IAT has several idiosyncratic features (44, 45). For example, it requires effortful stimulus categorization, and its proposed mechanism is response competition. Do our findings generalize to other measures of automatic evaluation?

In Experiment 3 we implemented the same learning phase as in Experiment 2 but measured automatic evaluations using an alternative assay: the Affect Misattribution Procedure (AMP) (46). Unlike the IAT, the AMP does not involve effortful stimulus categorization, and its proposed mechanism is the automatic activation of affect rather than response competition. Replicating the results of Experiment 2 with the AMP would strongly suggest that they reflect genuine contributions of model-based algorithms to automatic evaluations, rather than artifacts of a single measurement tool.

Model-Based Preferences Were Successfully Manipulated. Both revaluation conditions successfully shifted model-based preferences, as indexed by deliberate evaluations (Fig. 5). As in Experiment 2, participants maintained a significant deliberate preference for Energy Source A across all three conditions, t(808) = 13.65, P < 0.0001, Cohen's d = 0.48, BF_{10} = 7.51 × 10^{34}, but the strength of this preference varied by condition, F(2, 806) = 31.63, P < 0.0001, η^2 = 0.07, BF_{10} = 1.05 × 10^{11}. Specifically, participants showed a stronger preference for Energy Source A in the baseline learning condition (M = 29.25, SD = 36.10) compared to the motivation revaluation (M = 7.68, SD = 33.88), t(806) = 7.25, P < 0.0001, and transition revaluation conditions (M = 11.39, SD = 33.43), t(806) = 6.13, P < 0.0001. Deliberate evaluations did not differ significantly across the two revaluation conditions, t(806) = −1.19, P = 0.235.

Model-Based Algorithms Shaped Automatic Evaluations. Critically, automatic evaluations, as measured by the AMP, closely mirrored deliberate evaluations (Fig. 5). They differed across conditions, F(2, 829) = 23.81, P < 0.0001, η^2 = 0.05, BF_{10} = 8.26 × 10^7, favoring Energy Source A more in the baseline learning condition (M = 0.21, SD = 0.40) than the motivation revaluation (M = 0.03, SD = 0.37), t(829) = 5.77, P < 0.0001, and transition revaluation conditions (M = 0.03, SD = 0.34), t(829) = 5.94, P < 0.0001. The two revaluation conditions did not significantly differ, t(829) = 0.04, P = 0.971.

Unlike the deliberate preference for Energy Source A, the automatic preference revealed by the AMP was not merely attenuated in the revaluation conditions—it was eliminated entirely. All results remained the same when analyses were repeated only using data from participants who passed all manipulation checks.

Combined with the results of Experiment 2, these findings unequivocally suggest that automatic evaluations are subject to model-based influences. If anything, in Experiment 3, automatic evaluations appeared even more model-based than deliberate evaluations: While the revaluation manipulations merely attenuated the initial deliberate preference, they fully eliminated the automatic preference.

General Discussion

One of the most influential ideas in the psychological sciences is that behavior emerges from two types of processes: those that are automatic and those that are deliberate (4, 22, 47–50). This distinction is foundational to understanding phenomena ranging from reasoning (47, 50) to habit formation (4, 51–54) to social cognition (29, 55–57), and is thought to coincide with the algorithmic distinction between model-based and model-free reinforcement learning (2, 7). A major implication of this idea—one

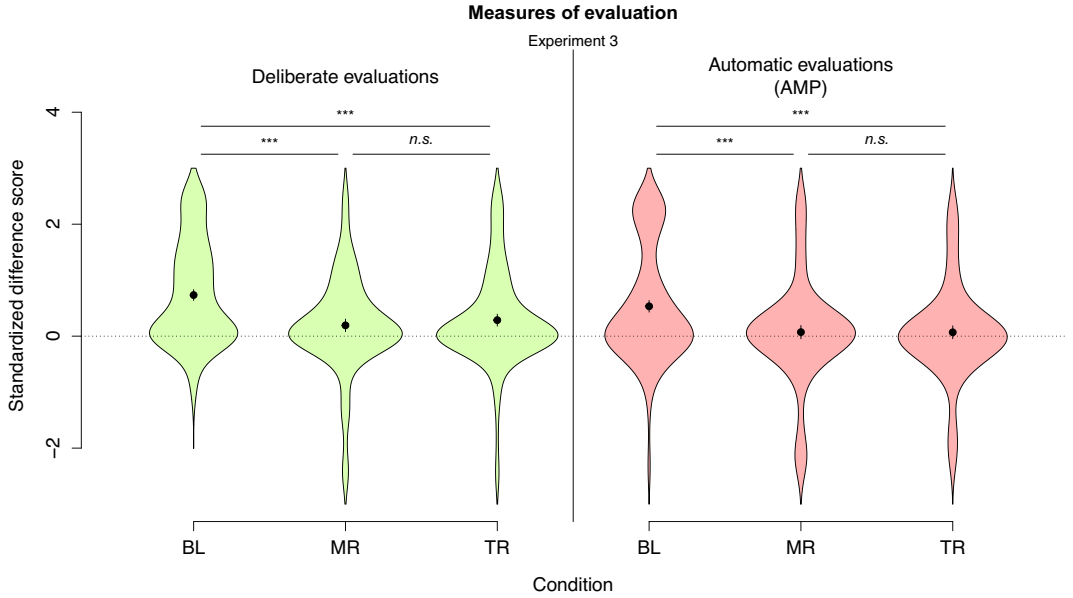

Fig. 5. Experiment 3 (n = 832): The distribution of deliberate evaluations and automatic evaluations (AMP scores) by learning condition. The x-axis shows condition (baseline learning, motivation revaluation, and transition revaluation), and the y-axis shows standardized difference scores between Energy Source A (which was initially rewarding) and Energy Source B (which was initially nonrewarding); positive scores indicate a preference for Energy Source A over Energy Source B. The solid dots denote condition means, and the error bars show 95% CIs. BL = baseline learning; MR = motivation revaluation; TR = transition revaluation; *n.s.* = not significant; ***P < 0.001.

that has driven decades of research and theorizing—is that model-based algorithms shape deliberate evaluations, but not automatic evaluations. The present findings suggest otherwise: Model-based algorithms appear to shape both deliberate and automatic evaluations.

Experiment 1 showed that the evidence against model-based input to automatic evaluation is less decisive than previously thought; confounding factors preserve the possibility that automatic evaluations are model-based. Experiments 2 to 3 eliminated these factors and found clear support for model-based contributions to automatic evaluation, which generalized across two widely used assays of automatic evaluative responding. In addition, when IAT data were analyzed using a multinomial processing tree approach, condition differences were shown to be driven entirely by automatic, rather than controlled, parameters, thus underscoring the finding that automatic evaluation is subject to model-based inputs.

This finding strains traditional accounts of when and why decision-makers err. Poor decisions are often blamed on automatic evaluative processing—a natural consequence of equating such processing with error-prone, model-free algorithms. By contesting this equivalence, the present work suggests that automatic evaluations may have surprisingly little to do with erroneous decision-making. People choose wrongly for many reasons (58–61), but automatic evaluative processing may not be among the primary culprits.

Other influential ideas may also need revising. For instance, when a psychological phenomenon relies primarily on model-based or model-free processing, it is often assumed to be controlled or automatic, respectively (9, 52, 57, 62). A case in point comes from research on moral behavior, where the process by which people learn to avoid harming others is thought to be primarily model-free. This finding has led to the conclusion that "[…] potentially harmful actions might be prioritized for automatic avoidance as a default"(57). However, the present work suggests that this conclusion may not be warranted—that the automaticity of a process cannot be reliably inferred from the algorithm that implements it. Similar implications extend to a range of phenomena where inferences about automaticity have been drawn from algorithmic insights, including the neural mechanisms of learning and decision-making (9), human development (52), and clinical disorders (62).

The present findings also contribute to the growing body of evidence suggesting that standard reinforcement learning tasks often make human decisions appear less model-based than they really are. Participants often misconstrue these tasks in ways that make their model-based decisions seem model-free (14, 27, 63). When disabused of their misconceptions, participants behave in ways that reveal a more extensive role for model-based algorithms in decision-making than traditionally understood. In line with this idea, the present findings suggest that automatic evaluations have been incorrectly classified as purely model-free due to participants' misinterpretation of the original task design. When this misinterpretation is addressed, it becomes clear that model-based algorithms extend to automatic evaluative processing, indicating an even broader impact than previously recognized.

The present findings also suggest that model-based influences could be underestimated for a related, but slightly different, reason (56, 64): Lab-based reinforcement learning tasks, much like Experiment 1 above, often rely on arbitrary sets of stimuli with arbitrary transitions between them. Such setups can create excessive working memory demands, which can interfere with the operation of model-based algorithms. By fostering conceptual scaffolding, introducing more naturalistic stimuli can reveal model-based influences even if the abstract structure of the task remains the same, as it did across the present Experiments 2 to 3 versus Experiment 1.

The present results also address a longstanding puzzle concerning how automatic evaluations change. While classic theories portray automatic evaluations as stable (65–67), a growing body of work suggests otherwise, showing that automatic evaluations can change completely in response to a single piece of new evidence (68–71). For example, in a paradigmatic study (68), participants read a single statement indicating that someone they believed was moral was, in fact, immoral. Automatic evaluations of the target person instantly inverted from positive to negative.

This finding, and others like it (70–75), show that automatic evaluations can swiftly change. However, the algorithms underlying this flexibility have not been established. They could, in principle, be solely model-free. Learning that someone is immoral, for instance, may simply change which model-free representation is retrieved when that person is encountered. The present work supports a different conclusion: Automatic evaluations owe their flexibility in part to model-based computations.

How do model-based algorithms shape automatic evaluations in the first place? The most straightforward hypothesis is that model-based algorithms can operate automatically, and thus can directly implement automatic evaluative processing. There is, however, another possibility. Research has revealed similarities between human reinforcement learning and a computational architecture called DYNA (76), in which model-based algorithms play a strictly indirect role in evaluative processing (15, 17–19). In DYNA, all evaluations—whether automatic or deliberate—are based solely on model-free value signals; model-based algorithms influence evaluations indirectly by shaping model-free value signals offline. Specifically, model-based algorithms are used to simulate trajectories through the environment; these trajectories are then fed into a model-free system, which computes new value signals as though the trajectories had been experienced first-hand.

DYNA predicts particular patterns of decision-making errors that have been observed in human behavior. It also predicts that model-based algorithms should shape automatic evaluation—a prediction that prior work appeared to rule out. The present work vindicates this prediction, demonstrating that model-based algorithms can indeed shape automatic evaluations. In turn, DYNA offers a compelling mechanism—offline simulation—that may explain how the present findings arise, although this possibility remains to be tested.

However it emerges, the link between model-based algorithms and automatic evaluations uncovered here sheds surprising light on how decision-makers evaluate their options. Counterintuitive though it may seem, model-based algorithms shape not only the evaluations that decision-makers deliberately generate, but also those that spring to mind spontaneously.

Materials and Methods

Institutional Approval and Informed Consent. The project was granted ethical approval by the Yale University Institutional Review Board (protocol #2000029368, "Social Learning and Memory") and the University of Illinois Urbana-Champaign Office for Protection of Research Subjects (protocol # IRB24-0118, "Measuring and shifting implicit attitudes"). Participants in all three experiments provided informed consent.

Open Science Practices. We report all manipulations, measures, and participant exclusions. The experiments were preregistered at https://aspredicted.org/74tc-ww88.pdf for Experiment 1, at https://aspredicted.org/5c2g-r3jn.pdf for Experiment 2, and at https://aspredicted.org/d6br-38s7.pdf for Experiment 3. All materials, raw data (including trial-level IAT data), and analysis scripts are available for download from the Open Science Framework (https://osf.io/c2zys/).

Participants. Participants were adult volunteers recruited from the Project Implicit educational website (https://implicit.harvard.edu/implicit/) who completed the experiments entirely online. 819 participants were recruited to participate in Experiment 1; the final sample size following preregistered participant exclusions was 751 (508 women, 218 men, 13 participants of other genders, M_{age} = 40 y, SD = 15 y). 1,013 participants were recruited to participate in Experiment 2; the final sample size following preregistered exclusions was 989 (655 women, 299 men, 24 participants of other genders, M_{age} = 38 y, SD = 16 y). 1,017 participants were recruited to participate in Experiment 3; the final sample size following preregistered exclusions was 832 (545 women, 242 men, 23 participants of other genders, M_{age} = 35 y, SD = 14 y). Further information on exclusion criteria and demographic variables is provided in *SI Appendix*.

Analytic Strategy. Statistical analyses were performed in the R statistical computing environment (version 4.4.2) using RStudio (version 2024.09.1+394). No a priori power calculations were conducted. However, the final sample size in Experiment 1 provided adequate (0.80) power to detect a small difference of Cohen's d = 0.20 in a t test and in Experiments 2 to 3 an effect size of η^2 = 0.01 in a one-way ANOVA design with three groups. Analyses were conducted using standard linear models, with planned comparisons implemented using the emmeans package (77) and accompanying Bayes Factors calculated using the BayesFactor package (78). The quad model analyses for Experiment 2 were conducted using the TreeBUGS package (79), with model specifications adapted from (80).

Dependent Measures. Following the learning phase of Experiment 1, participants first completed measures of a) transition structure (4 items) and b) choice (4 items), as well as c) a standard 5-block Implicit Association Test (IAT) as a measure of automatic evaluations of the stage 1 stimuli (in the same fixed order). The transition structure and forced choice measures were structurally identical to the choice trials from the learning phase but involved no feedback. Each participant completed four transition structure and four forced choice trials following part 1 of the learning phase, and participants in the transition revaluation condition completed four additional trials of each kind following the revaluation manipulation. Following the IAT, participants completed two self-report measures of deliberate evaluation and four self-report measures of moral character judgments, in the same fixed order.

In Experiment 2, participants completed a standard 5-block IAT as a measure of automatic evaluations of the stage 1 stimuli, followed by self-report measures of a) deliberate evaluation (2 items) and b) transition structure (5 items). Given the focus of the present work on automatic evaluation, the IAT was always administered first. Experiment 3 followed the same procedure as Experiment 2, with the exception that automatic evaluations were measured via the Affect Misattribution Procedure (AMP) rather than the IAT. Additional information on the dependent measures in all three experiments, including the verbatim text of all items, is reported in the *SI Appendix*.

Automatic evaluations. The Implicit Association Test (IAT) used to measure automatic evaluations of Laapians and Niffians in both conditions of Experiment 1 consisted of five blocks. Participants used the E and I keys to categorize stimuli on the IAT.

In the first block (category practice), participants made 20 speeded single categorization judgments of Laapian and Niffian stimuli. The category labels were *Laapians* and *Niffians*. The Laapian stimuli included *Caalap*, *Feelslap*, *Gabeelap*, *Ineelap*, and *Maasolap*; the Niffian stimuli included *Ibbonif*, *Jabbunif*, *Lebbunif*, *Mettanif*, and *Oballnif*. In the second block (attribute practice), participants made 20 speeded single categorization judgments of good and bad stimuli. The attribute labels were *Good* and *Bad*. The good stimuli included *Love*, *Peace*, *Joy*, *Happy*, *Sweet*, *Glory*, and *Lucky*; the bad stimuli included *Hate*, *War*, *Devil*, *Bomb*, *Bitter*, *Agony*, and *Grief*. In the third block (first combined block), participants made 40 speeded dual categorization judgments on which Laapians and good stimuli were assigned to one response key and Niffians and bad stimuli were assigned to the other response key (or the opposite, depending on counterbalancing; see below). In the fourth block (second category practice), participants made 20 speeded single categorization judgments of Laapian and Niffian stimuli, with the assignment to response keys reversed compared to the first block. Finally, in the fifth block (second combined block), participants made 40 speeded dual categorization judgments on which the mapping of categories to attributes was reversed compared to the third block.

The order of the two combined blocks was randomized such that some participants completed the Laapian/good (Niffian/bad) block first and the Niffian/good (Laapian/bad) block second, whereas for other participants the order was reversed. The IAT was scored using the improved scoring algorithm (81) such that a positive score indicates a difference in favor of the initially rewarding over the initially nonrewarding group.

In Experiment 2, the IAT followed the same procedure, but the category labels were *Pinecones* and *Seeds*, and category stimuli were images of pinecones and seeds. Moreover, the words *Good*, *Great*, *Fantastic*, *Pleasant*, and *Wonderful* were

used as good attribute stimuli; and the words *Awful*, *Bad*, *Horrible*, *Terrible*, and *Unpleasant* were used as bad attribute stimuli.

In Experiment 3, the Affect Misattribution Procedure (AMP) (46) was used as a measure of automatic evaluations. The AMP consisted of 40 trials. Each AMP trial involved the presentation of a prime stimulus for 75 ms, a blank screen for 125 ms, and a target stimulus for 100 ms, followed by a noise mask until the participant responded using the E key ("unpleasant") or I key ("pleasant") on their keyboard. The same images of pinecones and seeds from Experiment 2 served as prime stimuli and abstract paintings from (82) served as target stimuli.

Participants were asked to evaluate the target stimuli while ignoring the primes; however, given the spatiotemporal proximity between primes and targets and the evaluatively ambiguous nature of the latter, evaluations of the targets can be used as an indirect index of the evaluation of the primes. The proportion of "pleasant" responses following initially nonrewarding stimulus primes was subtracted from the proportion of "pleasant" responses following initially rewarding stimulus primes to calculate AMP scores.

Deliberate evaluations. In Experiment 1, the measures of deliberate evaluation asked participants to indicate, using a 7-point scale, how warmly or coldly they felt toward each social group. The response options were labeled "Extremely warmly," "Warmly," "Somewhat warmly," "Neither warmly nor coldly," "Somewhat coldly," "Coldly," and "Extremely coldly." A difference score was created by subtracting the deliberate evaluation of the initially nonrewarding group from the deliberate evaluation of the initially rewarding group.

In Experiments 2 to 3, the measures of deliberate evaluation asked participant to indicate, using a 100-point sliding scale, how warmly or coldly they felt toward each resource (pinecones vs. seeds). These two measures were administered in randomized order. A difference score was created by subtracting the deliberate evaluation of the initially nonrewarding stimulus from the deliberate evaluation of the initially rewarding stimulus.

Morality Judgments. In Experiment 1, the self-report measures of moral character asked participants to indicate, using 9-point scales, to what extent each social group was good and moral vs. bad and immoral. The response options were labeled "Completely disagree," "Strongly disagree," "Moderately disagree," "Slightly disagree," "Neither agree nor disagree," "Slightly agree," "Moderately agree," "Strongly agree," and "Completely agree." A double difference score was created by first subtracting the immoral response from the moral response for each social group and then taking the difference of those two difference scores such that higher scores indicate a preference for the initially rewarding over the initially nonrewarding group.

Data, Materials, and Software Availability. Data, analysis code, and materials have been deposited in the Open Science Framework (https://doi.org/10.17605/OSF.IO/C2ZYS) (83).

1. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, ed. 2, 2018).
2. N. Drummond, Y. Niv, Model-based decision making and model-free learning. *Curr. Biol.* **30**, R860–R865 (2020).
3. N. D. Daw, P. Dayan, The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130478 (2014).
4. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
5. S. J. Gershman, K. A. Norman, Y. Niv, Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci.* **5**, 43–50 (2015).
6. S. J. Gershman, D. M. Blei, Y. Niv, Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
7. N. D. Daw, J. P. O'Doherty "Multiple systems for value learning" in *Neuroeconomics*, P. W. Glimcher, E. Fehr, Eds. (Academic Press, ed. 2, 2014), pp. 393–410.
8. H. J. Don, M. B. Goldwater, A. R. Otto, E. J. Livesey, Rule abstraction, model-based choice, and cognitive reflection. *Psychon. Bull. Rev.* **23**, 1615–1623 (2016).
9. Y. Huang, Z. A. Yaple, R. Yu, Goal-oriented and habitual decisions: Neural signatures of model-based and model-free learning. *NeuroImage* **215**, 116834 (2020).
10. W. Kool, S. J. Gershman, F. A. Cushman, Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol. Sci.* **28**, 1321–1333 (2017).
11. B. Kurdi, S. J. Gershman, M. R. Banaji, Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proc. Natl. Acad. Sci.* **116**, 6035–6044 (2019).
12. A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, N. D. Daw, Working-memory capacity protects model-based learning from stress. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20941–20946 (2013).
13. A. R. Otto, A. Skatova, S. Madlon-Kay, N. D. Daw, Cognitive control predicts use of model-based reinforcement learning. *J. Cogn. Neurosci.* **27**, 319–333 (2015).
14. C. Feher da Silva, G. Lombardi, M. Edelson, T. A. Hare, Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nat. Hum. Behav.* **7**, 956–969 (2023).
15. S. J. Gershman, A. B. Markman, A. R. Otto, Retrospective revaluation in sequential decision making: A tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182–194 (2014).
16. W. Kool, F. A. Cushman, S. J. Gershman "Chapter 7—Competition and cooperation between multiple reinforcement learning systems" in *Goal-Directed Decision Making*, R. Morris, A. Bornstein, A. Shenhav, Eds. (Academic Press, 2018), pp. 153–178.
17. I. Momennejad, Learning structures: Predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* **32**, 155–166 (2020).
18. I. Momennejad *et al.*, The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
19. I. Momennejad, A. R. Otto, N. D. Daw, K. A. Norman, Offline replay supports planning in human reinforcement learning. *eLife* **7**, e32548 (2018).
20. A. R. Otto, S. J. Gershman, A. B. Markman, N. D. Daw, The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761 (2013).
21. G. Keren, Y. Schul, Two is not always better than one: A critical evaluation of two-system theories. *Perspect. Psychol. Sci.* **4**, 533–550 (2009).
22. D. E. Melnikoff, J. A. Bargh, The mythical number two. *Trends Cogn. Sci.* **22**, 280–293 (2018).
23. A. Moors, J. De Houwer, Automaticity: A theoretical and conceptual analysis. *Psychol. Bull.* **132**, 297–326 (2006).
24. J. R. Stroop, Studies of interference in serial verbal reactions. *J. Exp. Psychol.* **18**, 643–662 (1935).
25. S.-Y. Kim, M.-S. Kim, M. M. Chun, Concurrent working memory load can reduce distraction. *Proc. Natl. Acad. Sci.* **102**, 16524–16529 (2005).
26. A. G. Greenwald, D. E. McGhee, J. L. K. Schwartz, Measuring individual differences in implicit cognition: The implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).
27. C. Feher da Silva, T. A. Hare, Humans primarily use model-based inference in the two-stage task. *Nat. Hum. Behav.* **4**, 1053–1066 (2020).
28. F. Cushman, Deconstructing intent to reconstruct morality. *Curr. Opin. Psychol.* **6**, 97–103 (2015).
29. L. M. Hackel, B. B. Doll, D. M. Amodio, Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nat. Neurosci.* **18**, 1233–1235 (2015).
30. L. M. Hackel, P. Mende-Siedlecki, D. M. Amodio, Reinforcement learning in social interaction: The distinguishing role of trait inference. *J. Exp. Soc. Psychol.* **88**, 103948 (2020).
31. F. Heider, *The Psychology of Interpersonal Relations* (John Wiley & Sons Inc, 1958).
32. B. Kurdi, A. R. Krosch, M. J. Ferguson, Implicit evaluations of moral agents reflect intent and outcome. *J. Exp. Soc. Psychol.* **90**, 103990 (2020).
33. B. F. Malle, How people explain behavior: A new theoretical framework. *Pers. Soc. Psychol. Rev.* **3**, 23–48 (1999).
34. N. Vélez, H. Gweon, Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Curr. Opin. Behav. Sci.* **38**, 110–115 (2021).
35. L. Winter, J. S. Uleman, When are social judgments made? Evidence for the spontaneousness of trait inferences *J. Pers. Soc. Psychol.* **47**, 237–252 (1984).
36. S. T. Fiske, A. J. C. Cuddy, P. Glick, Universal dimensions of social cognition: Warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
37. G. P. Goodwin, Moral character in person perception. *Curr. Dir. Psychol. Sci.* **24**, 38–44 (2015).
38. G. P. Goodwin, J. Piazza, P. Rozin, Moral character predominates in person perception and evaluation. *J. Pers. Soc. Psychol.* **106**, 148–168 (2014).
39. P. Dayan, Improving generalization for temporal difference learning: The successor representation. *Neural Comput.* **5**, 613–624 (1993).
40. K. C. Klauer, Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika* **71**, 7–31 (2006).
41. K. C. Klauer, Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika* **75**, 70–98 (2010).
42. F. R. Conrey, J. W. Sherman, B. Gawronski, K. Hugenberg, C. J. Groom, Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *J. Pers. Soc. Psychol.* **89**, 469–487 (2005).
43. M. Hütter, K. C. Klauer, Applying processing trees in social psychology. *Eur. Rev. Soc. Psychol.* **27**, 116–159 (2016).
44. J. De Houwer, A. Moors, "Implicit measures: Similarities and differences" in *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications* (The Guilford Press, 2010), pp. 176–193.
45. J. De Houwer, S. Teige-Mocigemba, A. Spruyt, A. Moors, Implicit measures: A normative analysis and review. *Psychol. Bull.* **135**, 347–368 (2009).
46. B. K. Payne, C. M. Cheng, O. Govorun, B. D. Stewart, An inkblot for attitudes: Affect misattribution as implicit measurement. *J. Pers. Soc. Psychol.* **89**, 277–293 (2005).
47. D. Kahneman, Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475 (2003).
48. M. I. Posner, C. R. R. Snyder, *Attention and Cognitive Control* (Psychology Press, 2004).
49. R. M. Shiffrin, W. Schneider, Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.* **84**, 127–190 (1977).
50. K. E. Stanovich, R. F. West, Individual differences in reasoning: Implications for the rationality debate? *Behav. Brain Sci.* **23**, 645–665 (2000).
51. B. W. Balleine, J. P. O'Doherty, Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69 (2010).
52. J. H. Decker, A. R. Otto, N. D. Daw, C. A. Hartley, From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychol. Sci.* **27**, 848–858 (2016).
53. A. Dickinson, Actions and habits: The development of behavioural autonomy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **308**, 67–78, (1985).
54. C. M. Gillan, A. R. Otto, E. A. Phelps, N. D. Daw, Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).
55. L. M. Hackel, J. J. Berg, B. R. Lindström, D. M. Amodio, Model-based and model-free social cognition: Investigating the role of habit in social attitude formation and choice. *Front. Psychol.* **10**, 2592 (2019).
56. L. M. Hackel, D. A. Kalkstein, P. Mende-Siedlecki, Simplifying social learning. *Trends Cogn. Sci.* **28**, 428–440 (2024).

57. P. L. Lockwood, M. C. Klein-Flügge, A. Abdurahman, M. J. Crockett, Model-free decision making is prioritized when learning to avoid harming others. *Proc. Natl. Acad. Sci.* **117**, 27719–27730 (2020).
58. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
59. K. J. Miller, A. Shenhav, E. A. Ludvig, Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
60. W. A. Wickelgren, Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol. (Amst.).* **41**, 67–85 (1977).
61. W. Wood, D. Rünger, Psychology of habit. *Annu. Rev. Psychol.* **67**, 289–314 (2016).
62. K. Foerde *et al.*, Deficient goal-directed control in a population characterized by extreme goal pursuit. *J. Cogn. Neurosci.* **33**, 463–481 (2021).
63. E. K. Buabang *et al.*, A goal-directed account of action slips: The reliance on old contingencies. *J. Exp. Psychol. Gen.* **152**, 496–508 (2023).
64. L. M. Hackel, D. A. Kalkstein, Social concepts simplify complex reinforcement learning. *Psychol. Sci.* **34**, 968–983 (2023).
65. B. Gawronski, G. V. Bodenhausen, Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychol. Bull.* **132**, 692–731 (2006).
66. A. P. Gregg, B. Seibt, M. R. Banaji, Easier done than undone: Asymmetry in the malleability of implicit preferences. *J. Pers. Soc. Psychol.* **90**, 1–20 (2006).
67. R. J. Rydell, A. R. McConnell, Understanding implicit and explicit attitude change: A systems of reasoning analysis. *J. Pers. Soc. Psychol.* **91**, 995–1008 (2006).
68. J. Cone, M. J. Ferguson, He did what? The role of diagnosticity in revising implicit evaluations *J. Pers. Soc. Psychol.* **108**, 37–57 (2015).
69. M. J. Ferguson, T. C. Mann, J. Cone, X. Shen, When and how implicit first impressions can be updated. *Curr. Dir. Psychol. Sci.* **28**, 331–336 (2019).
70. B. Kurdi, M. R. Banaji, Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *J. Exp. Psychol. Gen.* **146**, 194–213 (2017).
71. D. E. Melnikoff, A. H. Bailey, Preferences for moral vs. immoral traits in others are conditional. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E592–E600 (2018).
72. J. De Houwer, P. Van Dessel, T. Moran, "Chapter Three–Attitudes beyond associations: On the role of propositional representations in stimulus evaluation" in *Advances in Experimental Social Psychology*, B. Gawronski, Ed. (Academic Press, 2020), pp. 127–183.
73. P. Van Dessel, J. De Houwer, A. Gast, C. T. Smith, M. De Schryver, Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *J. Exp. Soc. Psychol.* **63**, 1–9 (2016).
74. P. Van Dessel, J. De Houwer, A. Gast, C. Tucker Smith, Instruction-based approach-avoidance effects. *Exp. Psychol.* **62**, 161–169 (2015).
75. D. E. Melnikoff, R. Lambert, J. A. Bargh, Attitudes as prepared reflexes. *J. Exp. Soc. Psychol.* **88**, 103950 (2020).
76. R. S. Sutton, Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bull.* **2**, 160–163 (1991).
77. R. V. Lenth, emmeans: Estimated marginal means, aka least-squares means. https://doi.org/10.32614/CRAN.package.emmeans. Deposited 20 October 2017.
78. R. D. Morey, J. N. Rouder, BayesFactor: Computation of bayes factors for common designs. https://doi.org/10.32614/CRAN.package.BayesFactor. Deposited 11 October 2012.
79. D. W. Heck, N. R. Arnold, D. Arnold, TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behav. Res. Methods.* **50**, 264–284 (2018).
80. C. T. Smith, J. Calanchini, S. Hughes, P. Van Dessel, J. De Houwer, The impact of instruction- and experience-based evaluative learning on IAT performance: A Quad model perspective. *Cogn. Emot.* **34**, 21–41 (2020).
81. A. G. Greenwald, B. A. Nosek, M. R. Banaji, Understanding and using the implicit association test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* **85**, 197–216 (2003).
82. J. H. Katz, T. C. Mann, X. Shen, J. A. Goncalo, M. J. Ferguson, Implicit impressions of creative people: Creativity evaluation in a stigmatized domain. *Organ. Behav. Hum. Decis. Process.* **169**, 104116 (2022).
83. D. E. Melnikoff, B. Kurdi, Data, analysis code, and materials for "Model-based algorithms shape automatic evaluative processing." Open Science Framework. https://doi.org/10.17605/OSF.IO/C2ZYS. Deposited 12 December 2024.