Clarifying the role of an unavailable distractor in human multiattribute choice

Yinan Cao[1] and Konstantinos Tsetsos[1,2]

[1] Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[2] School of Psychological Science, University of Bristol, Bristol, United Kingdom

Correspondence: ycaoneuro@gmail.com, k.tsetsos@bristol.ac.uk

Abstract

Decisions between two economic goods can be swayed by a third *unavailable* 'decoy' alternative, which does not compete for choice, notoriously violating the principles of rational choice theory. Although decoy effects typically depend on the decoy's position in a multiattribute choice space, recent studies using risky prospects (i.e., varying in reward and probability) reported a novel 'positive' decoy effect operating on a single 'value' dimension: the higher the 'expected value' of an unavailable (distractor) prospect was, the easier the discrimination between two available target prospects became, especially when their expected-value difference was small. Here we show that this unidimensional distractor effect affords alternative interpretations: it occurred because the distractor's expected value covaried positively with the subjective utility difference between the two targets. Looking beyond this covariation, we report a modest 'negative' distractor effect operating on subjective utility, as well as classic multiattribute decoy effects. A normatively meaningful model (selective integration), in which subjective utilities are shaped by 'intra-attribute' information distortion, reproduces the multiattribute decoy effects, and as an epiphenomenon, the negative unidimensional distractor effect. These findings clarify the modulatory role of an unavailable distracting option, shedding fresh light on the mechanisms that govern multiattribute decisions.

Introduction

Humans strive to make good decisions but are rarely veridical judges of the world. Instead, our judgments are swayed by seemingly irrelevant factors[1] and our preferences are improvised as we go along[2]. For instance, echoing well-documented optical illusions, we may perceive the same doughnut as larger on a tiny plate than on a bigger one (**Fig. 1a**). Analogous distortions are encountered when we choose among alternatives that vary in more than one dimension or *attribute*. For instance, in the attraction effect[3], adding a similar but less advantageous 'decoy' option (a disk drive with a good storage capacity but at a high price) can make a target alternative (a lower-priced disk with more storage) appear more appealing than a competing alternative (an affordable disk with lower storage).

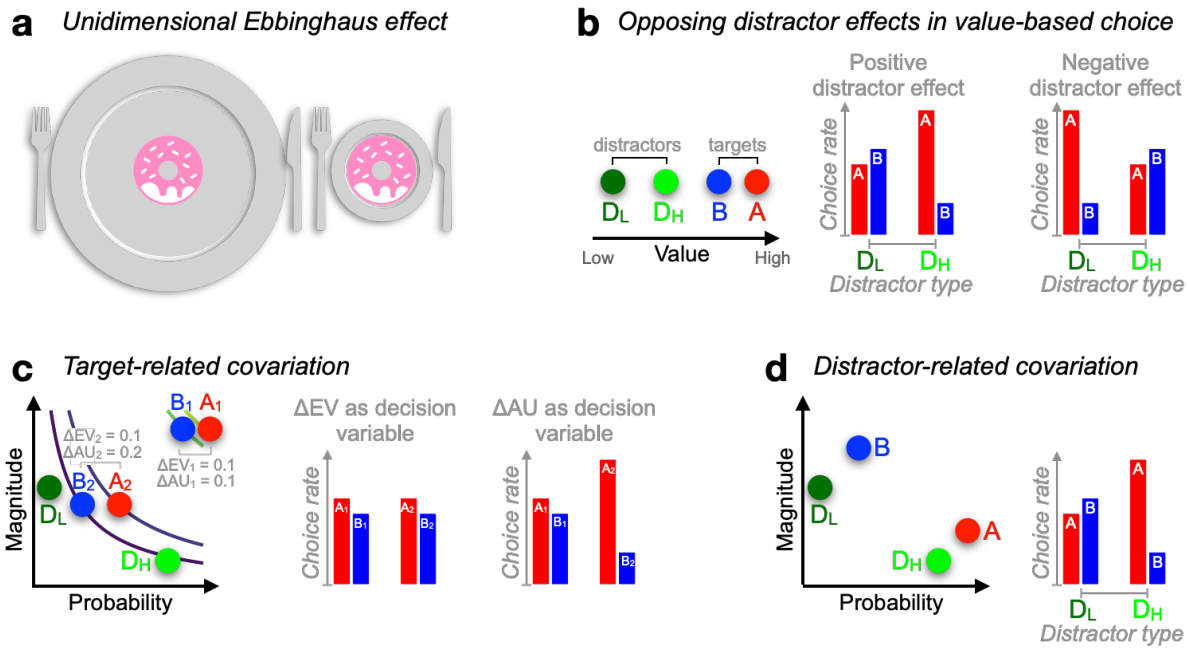


Figure 1. Context effects in decision-making. (**a**) Judgment about object size can be influenced by the surround. (**b**) Documented positive[4] or negative[5] distractor effects occur at the level of unidimensional value. A higher-valued distractor (D_H), compared with a lower-valued distractor (D_L), can *increase* the choice rate of A relative to B (*positive* distractor effect) but can also *decrease* the choice rate of A relative to B (*negative* distractor effect), leading to opposing distractor effects in past reports[4,5]. (**c**) During risky choices, if people derive subjective utilities of prospects by integrating reward magnitude and probability additively (additive utility or AU), and if targets $\{A_2, B_2\}$ are more frequently paired with distractor D_H while targets $\{A_1, B_1\}$ with D_L, then a positive distractor effect could be explained by additive integration (because $\Delta AU_1 < \Delta AU_2$) rather than the distractor's expected value (EV: magnitude x probability). Curves: EV indifference lines. (**d**) Similarly, relative preference changes for target alternatives A and B can be driven by an *attraction-effect* bias[3,6] that depends on the position of the decoy D with respect to each target alternative. If D_H appears more often closer to A while D_L being more often closer to B, a positive distractor effect will ensue as a by-product of the attraction effect. In all illustrations, A-alternatives are always better than B-alternatives, with D-alternatives denoting the distractor (decoy) alternatives.

The attraction effect together with related contextual *preference-reversal* phenomena violate the very notion of rational-choice explanation according to which preferences should be menu-invariant[7,8]. Instead, these phenomena suggest that preferences are menu-variant or *context-*

sensitive, with the subjective utility (a notion of attractiveness) of an economic alternative being malleable to the properties of other available alternatives. Preference reversals have been typically reported in multiattribute choice settings, instigating a departure from 'value-first' utility theories[9]—in which attributes are integrated within each option independently—towards the development of novel theories of multiattribute choice[10–14].

In contrast to multiattribute decisions, decisions over alternatives varying along a single attribute have been typically regarded as complying with the principles of rational choice theory[15]. A plausible account for this dichotomy could be that multiattribute decisions are more complex. Recently however, a new type of context effect has been reported in unidimensional multialternative decisions. In the so-called *negative distractor* effect, humans become more indifferent between two unequal alternatives of high reward value when a third distracting option has a high, as opposed to a low, reward value[5] (**Fig. 1b**; although see a recent replication failure of this effect[16]). This negative distractor effect is theoretically important because it shows that violations of rationality may not be restricted to complex multiattribute decisions[17]. In turn, this motivates the development of a cross-domain, unified theory of decision irrationality.

Interestingly, in a series of more recent experiments involving risky prospects (alternatives offering reward outcomes with some probabilities), the opposite pattern was predominantly documented[4,18] (**Fig. 1b**): the decision accuracy of choosing the best of two target prospects is particularly compromised by a low expected-value, 'unavailable' prospect (distractor) that should have never been in contention. This disruption effect fades away as the distractor's expected value increases, leading to a *positive* distractor effect. However, Gluth and colleagues (2018) casted doubts on this positive distractor effect, claiming that it arises due to statistical artefacts[19]. Later, in return, Chau et al. (2020) defended the positive distractor effect by emphasising its robust presence in difficult trials only (it tends to reverse into a negative distractor effect in easy trials, i.e., distractor effects *interact* across different parts of the decision space[18]).

Overall, the abovementioned risky-choice studies paint a rather complex empirical picture on the ways a distractor alternative impacts decision quality. Notably, in these studies, key design aspects and analyses rested on the assumption that participants' decisions were fully guided by the expected value of the prospects. However, since Bernoulli's seminal work, it has been well established that expected value is a theoretical construct that often fails to adequately describe human decisions under uncertainty[20,21]. Instead, human decisions under uncertainty are guided by subjective utilities that are shaped by a plethora of non-normative factors[22]. Such factors—including the absolute reward magnitude of a prospect[23], the riskiness of a prospect[24], or even the sum of the normalised reward and probability[25–27]—can perturb people's valuations of risky prospects in ways that the expected-value framework does not capture[28]. We reasoned that if such factors covaried with the expected value of the distractor, then the distractor effects reported previously could, partially or fully, afford alternative interpretations.

It could be argued that if the experimental choice-sets are generated randomly and afresh for each participant, such covariations between the distractor's expected value and other factors are likely to be averaged out. However, we note that all previous studies reporting positive and

interacting distractor effects[4,18] used the exact same set of trials for all participants. This set of trials was generated pseudo-randomly, by resampling choice-sets until the expected value of the distractor was sufficiently decorrelated from the expected-value difference (choice *difficulty*) of the two targets. On the positive side, presenting the same set of trials to all participants eliminates the impact of stimulus variability on the group-level behaviour[29]. On the negative side, using a single set of trials in conjunction with this decorrelation approach, increases the risk of introducing unintended confounding covariations in the elected set of trials, which in turn will have consistent impact on the group-level behaviour.

Here we outline two classes of unintended covariations that could potentially explain away distractor effects in these datasets. First, the expected value (EV) of the distractor could potentially covary with specific (and influential to behaviour) reward/probability regularities in the two target prospects (we call them *target-related* covariations). For instance, if people valuate prospects by simply adding up their payoff and probability information (hereafter *additive utility* or *AU*)[25,27,30], two target prospects both offering £1 with probability of 0.9 vs. 0.8, respectively (i.e., EV difference: 0.1, AU difference: 0.1), will be less easily discriminable than two other prospects both offering £0.5 with probability of 0.5 vs. 0.3, respectively (i.e., EV difference: 0.1, AU difference: 0.2; **Fig. 1c**). Crucially, if the first choice-set (AU difference: 0.1) is more frequently associated with low expected-value distractors than the second one (AU difference: 0.2), then a positive distractor effect could be attributable to additive integration rather than the distractor's EV (**Fig. 1c**). Second, the EV of the distractor alternative could covary with its relative position in the reward-probability space (we call these *distractor-related* covariations). This would influence decision accuracy because, as outlined earlier, the relative position of a decoy alternative in the multiattribute choice space can induce strong choice biases[12,13]. For illustration purposes only, we assume that the distractor boosts the tendency of choosing a nearby target akin to the attraction effect[3,6] (**Fig. 1d**). Under this assumption, if distractors with high expected values happen to appear closer to the correct target (i.e., the target with the highest expected value), a positive distractor effect could be entirely attributable to the attraction effect.

The aim of this paper is to re-assess distractor effects in the relevant, previously published datasets[4,18,19] while looking beyond these two classes of potentially confounding covariations (target-related and distractor-related). We began with establishing that the first class of target-related covariations is indeed present in the examined datasets, with positive and interacting 'notional' distractor effects being evident even in matched *binary* trials, in which the distractor alternative was not present. Using a novel baselining approach, we asked if there are residual distractor effects when the influence of these unintended covariations is controlled for, reporting that distractor effects are eradicated. We then pinpointed these target-related confounding covariations to people's strong tendency to integrate reward and probability additively, and not multiplicatively. Moving forward, defining the key target and distractor *utility* variables, not using expected value, but subjective (additive) utility, revealed a modest negative distractor effect. Moving on to the second class of distractor-related covariations, we established that choice accuracy was lawfully influenced by the position of the distractor in the multiattribute space (**Fig. 1d**), consistent with a large body of empirical work in multiattribute choice[6,12]. This 'decoy' influence was most pronounced when the distractor alternative was close to the high-EV (correct) target in the multiattribute space, yielding an attraction effect (i.e., a boost in accuracy) when the distractor was inferior, and a repulsion effect (i.e., a

reduction in accuracy) when the distractor was superior (in subjective utility) to the target. Of note, this decoy influence peaking around the high-EV target, essentially re-describes a negative distractor effect without evoking ad-hoc 'unidimensional distractor' mechanisms other than those that are needed to produce classic multiattribute context effects[12,31].

Overall, our analyses update the state-of-the-art by suggesting that, when confounding covariations are controlled for, only a modest negative distractor effect survives. Further, it is conceivable that this distractor effect mirrors asymmetric classic multiattribute context effects, which occurred robustly in the examined datasets.

Results

We re-analysed 5 datasets from published studies (Chau et al.[4]; Gluth et al.[19]; *N* = 144 human participants) of a speeded multiattribute decision-making task. Participants were shown choice alternatives that varied along two distinct reward attributes: reward magnitude (X) and reward probability (P), which mapped to visual features (colourful slanted bars as options in **Fig. 2a**). After learning the feature-to-attribute mapping, participants were instructed to choose the most valuable one among multiple options placed in this multiattribute space (**Fig. 2a**) on each trial and to maximise the total reward across trials. In the ternary-choice condition, one option (the 'distractor', or *D*) was flagged *unavailable* for selection early in the trial. The expected value (or *EV*, i.e., X multiplied by P) of the higher- and lower-value *available* alternatives ('targets' *H* and *L*, respectively), and of the *unavailable* distractor were denoted by *HV*, *LV*, and *DV*, respectively.

Rational choice theory posits that an unavailable D should not in any way alter the relative choice between two available H and L targets. However, the behavioural data in this task[4] challenged this view. By examining the probability of H being chosen over L in ternary trials, i.e., the *relative* choice accuracy, Chau et al.[4,18] reported that a *relative* distractor variable 'DV – HV' (the EV difference between the distractor and the best available target) altered relative accuracy. We note that other recent studies, using different stimuli, quantified the distractor influence by means of an *absolute* distractor variable DV[5,16]. Wherever possible, we quantify distractor effects using both the relative (DV – HV) and absolute (DV) distractor variables.

To begin with, we note that, in addition to ternary trials in which D was present (but unavailable), participants encountered 'matched' binary trials in which only the H and L target alternatives were shown. These binary trials are ideally suited for assessing the extent to which previously reported distractor affects can be attributed to covariations between the distractor variable and target-related properties. This is because, for each ternary trial, we can derive a *respective* binary-choice baseline accuracy from binary trials that had the exact same targets but no distractor (**Figure 2—figure supplement 1c**). Given that participants never saw D in binary trials, the distractor variable is *notional* and should have no effect on the binary baseline accuracies. However, if D does paradoxically 'influence' binary accuracies, then this would signal that the distractor variable covaries with other unspecified modulators of choice accuracy. We dub any effect that D has upon binary-choice accuracy as the 'notional distractor effect'. We emphasise here that a notional distractor effect is not a genuine empirical phenomenon but a tool to diagnose target-/distractor-related covariations in the experimental design.

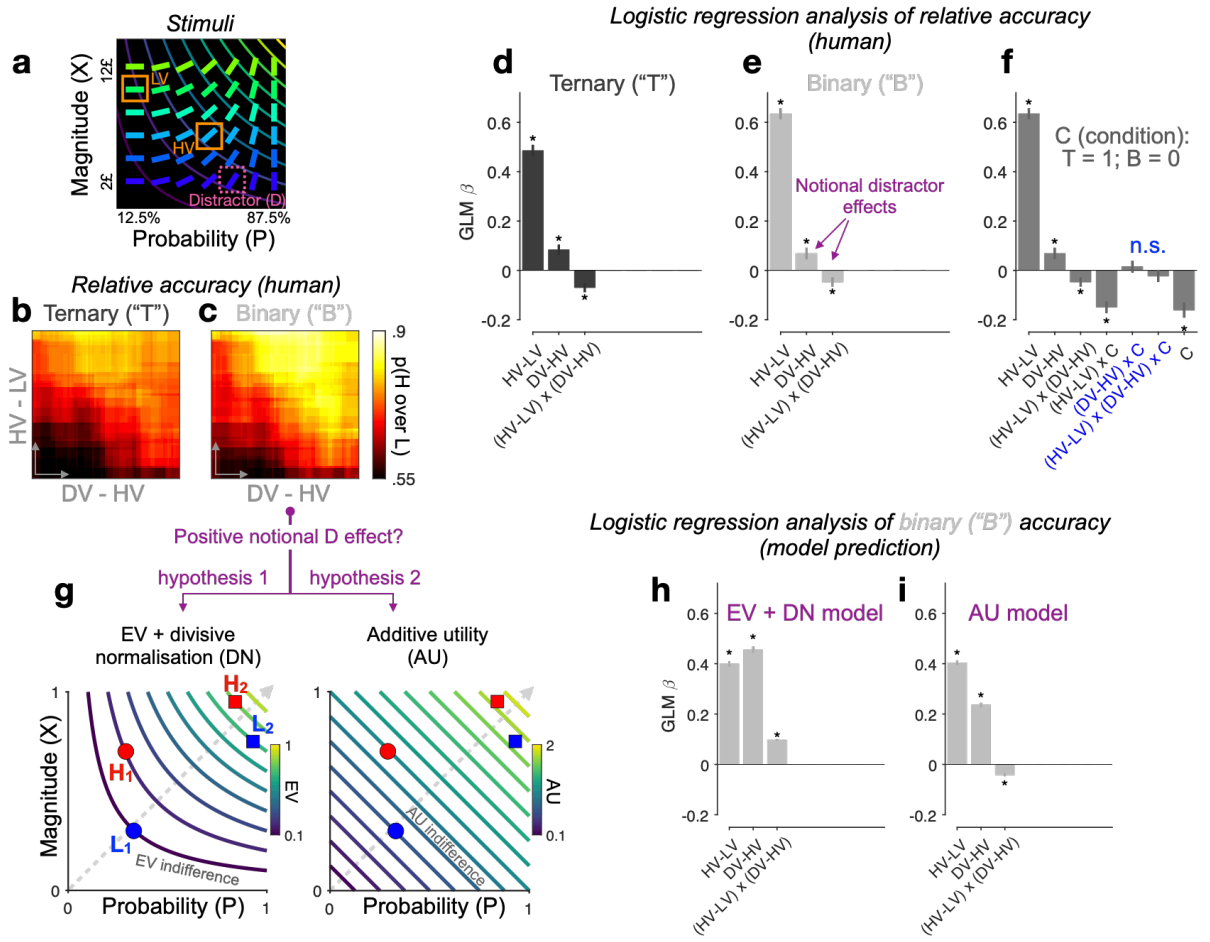


Figure 2. Re-assessing distractor effects beyond target-related covariations in ternary- and binary-choice trials. (**a**) Multiattribute stimuli in a previous study[4]. Participants made a speeded choice between two available options (HV: high expected value; LV: low expected value) in the presence (ternary) or absence (binary) of a 3rd distractor option (*'D'*). Three example stimuli are labelled for illustration purposes only. In the experiment, D was surrounded by a purple square to show that it should not be chosen. (**b**, **c**) Relative choice accuracy (probability of H choice among all H and L choices) in ternary trials (panel b) and in binary-choice baselines (panel c) plotted as a function of both the expected value (EV) difference between the two available options (HV – LV) (y-axis) and the EV difference between D and H (x-axis). Relative choice accuracy increases when HV – LV increases (bottom to top) as well as when DV – HV increases (left to right, i.e., *positive* D effect). (**d**, **e**, **f**) Predicting relative accuracy in human data using regression models (Methods). Asterisks: significant effects ($p < .05$; 2-sided 1-sample *t*-tests of GLM coefficients against 0) following Holm's sequential Bonferroni correction for multiple comparisons. (**g**) Rival hypotheses underlying the positive notional distractor effect on binary-choice accuracy. Left: EV indifference contour map; Right: additive-utility (AU) indifference contour map. Overall utility remains constant across all points on the same indifference curve and increases across different curves in evenly spaced steps along the direction of the dashed grey line. Decision accuracy scales with Δ(utility) between H and L in binary choices. Left (hypothesis 1): Δ(EV) = (HV - LV)/(HV + LV) by virtue of divisive normalisation; because $HV_2 - LV_2 = HV_1 - LV_1$, and $HV_2 + LV_2 > HV_1 + LV_1$, $\Delta(EV_2) < \Delta(EV_1)$. Right (hypothesis 2): $\Delta(AU) = AU_H - AU_L$; $\Delta(AU_2) < \Delta(AU_1)$. (**h**, **i**) Regression analysis of model-predicted accuracy in binary choice. EV + DN model corresponds to hypothesis 1 whilst AU model corresponds to hypothesis 2 in panel g. Error bars = ± s.e.m. (*N* = 144 participants).

Re-assessing distractor effects beyond target-related covariations

We used logistic regression (generalised linear model, or GLM) to quantify the effect of the distractor variable on relative choice accuracy (the probability of H choice among all H and L choices). Differently from previous studies, which focused primarily on ternary-choice trials, here we analysed two different dependent variables (**Fig. 2d—f**): ternary-choice relative accuracies ('T') and their *respective* baseline accuracies ('B') (see Methods and **Figure 2—figure supplement 1c**). For the ternary-choice condition, our GLM reveals a significant main effect of the relative distractor variable 'DV – HV' ($t(143) = 3.89$, $p < .001$), and a significant interaction between this distractor variable and the index of decision difficulty 'HV – LV' on relative choice accuracy, $t(143) = -3.67$, $p < .001$ (**Fig. 2d**). These results agree with previous reports when analysing ternary choice alone[4,18]. Turning into the matched binary baseline accuracies, the GLM coefficients bear a striking resemblance to those of the ternary-choice GLM (see **Fig. 2e**; also compare **Fig. 2b** with **2c** for a stark resemblance between T and B accuracy patterns), with (this time) *notional* positive and interacting distractor effects being observed. Crucially, neither the main distractor nor the (HV – LV) x (DV – HV) interaction variables are modulated by 'Condition' in a GLM combining T and B together, $|t(143)| < .92$, $p > .72$ (**Fig. 2f**). We see similar results when the relative distractor variable DV – HV is replaced by the absolute distractor variable 'DV' or when a covariate HV + LV is additionally included: the distractor's EV had no differential effect on relative accuracy across binary vs. ternary condition (**Figure 2—figure supplement 2, 3**). These results equate the distractor effects in ternary trials (D present) with the notional distractor effects in binary trials (D absent), indicating that the former arose for reasons other than the properties of D. Specifically, the notional distractor effects (**Fig. 2e**) in binary trials (with only H and L stimuli present) indicate that the value of D covaries with target-related properties (other than HV – LV, which was already a regressor) that modulate choice accuracy. Next, we use computational modelling to unveil these confounding target-related properties.

Integrating reward and probability information additively in risky choice

The original study reported a surprising phenomenon[4]: decisions seem to be particularly vulnerable to disruption by a 3rd low-value, as opposed to high-value, distracting alternative, i.e., a positive distractor effect. The very definition of this effect hinges upon the *a priori* assumption that decisions rely on calculating and subsequently comparing the expected values (EVs) across alternatives. An alternative prominent idea is that participants eschew EV calculations[32]; instead, they compute the difference between alternatives *within* each attribute and simply sum up these differences across attributes. As mentioned in the Introduction, we refer to this class of strategies that involve the additive (independent) contributions of reward and probability as additive utility (AU). Indeed, past studies in binary risky choice have reported decision strategies in humans and other animals[25,27,30] based on a weighted sum of the attribute differences, i.e., a decision variable equivalent to the additive utility (AU) difference between alternatives, $\Delta(AU) = \lambda(HX - LX) + (1 - \lambda)(HP - LP)$, where $0 \leq \lambda \leq 1$. Although the additive combination of reward and probability may not be generalisable in all types of risky choice tasks, it could viably govern decisions in the simple task illustrated in **Fig. 2a**. Of note, we came to notice that the key distractor variable DV – HV positively covaries with the Δ(AU) between H and L across all choice sets (e.g., Pearson's $r(148) = .314$, $p < .0001$, in the case of equal weighting between X and P, $\lambda = .5$). This unintended covariation arose in the original

study possibly due to a deliberate attempt to decorrelate two EV-based quantities (DV – HV and HV – LV) in the stimuli[4]. The correlation between DV – HV and Δ(AU) is stronger in more difficult trials ($r(71) = .52$, $p < .0001$, shared variance = 27%; splitting 150 trials into difficult vs. easy by the median of HV – LV[18]) than in easier trials ($r(59) = .43$, $p < .001$, shared variance = 19%), mirroring both the positive (overall positive correlation) and the interacting (change of correlation strength as a function of difficulty) notional distractor effects on binary choice.

Additive integration explains notional distractor effects in binary-choice trials

Notably, DV – HV also negatively covaries with HV + LV (Pearson's $r(148) = -.78$, $p < .001$), which potentially leads to another explanation of why binary choice accuracy seems lower as the matched DV – HV variable decreases. This explanation appeals to the divisive normalisation (DN) model based on EV[5,33]. Imagine choosing between two prospects, H and L, in two different choice-sets, H_1 vs. L_1, and H_2 vs. L_2, with the EV difference between H and L being the same across the two choice-sets (**Fig. 2g** left panel). DN applied to EVs ('EV + DN', hereafter) will shrink the EV difference between H and L more aggressively for set 2 than for set 1 because of the larger denominator in set 2, rendering set 2 a harder choice (lower accuracy). It is important to note that, in line with previous analyses[4,19], adding the HV + LV covariate to the GLMs does not explain away the interacting notional distractor effects in binary trials (**Figure 2—figure supplement 3**). Qualitatively, the above 2 hypotheses ('AU' vs. 'EV + DN') both predict a positive main notional distractor effect on binary choice accuracy, but their predictions for the (HV – LV) x (DV – HV) interaction could differ. For instance, DN might even predict a slightly positive, rather than negative, interacting distractor effect on binary accuracy—in this stimulus set, the divisively normalised Δ(EV) happens to be more positively correlated with DV – HV in easier trials (shared variance = 39.8%) than in harder trials (shared variance = 36.7%).

It is also important to note that AU on its own can approximate this EV sum effect by nearly tripling the utility difference between H and L in set 1 compared with that in set 2, which also renders set 2 a harder choice (**Fig. 2g** right panel). These two models can thus mimic each other. To better distinguish between these candidate hypotheses, we fit a model with a softmax decision rule to each participant's binary choice probabilities (Methods). As expected, both the EV + DN model and the AU model (with a free λ parameter; mean λ estimate: .46, SE: .017) predict a positive 'distractor' main effect on binary accuracy (**Fig. 2h**, **i**). However, the AU model, but not the EV + DN model, reproduces a negative (HV – LV) x (DV – HV) interaction effect on accuracy (**Fig. 2i**), mirroring the human data (**Fig. 2e**). The AU model thus qualitatively provides a parsimonious account of why notional distractor effects occurred in binary trials.

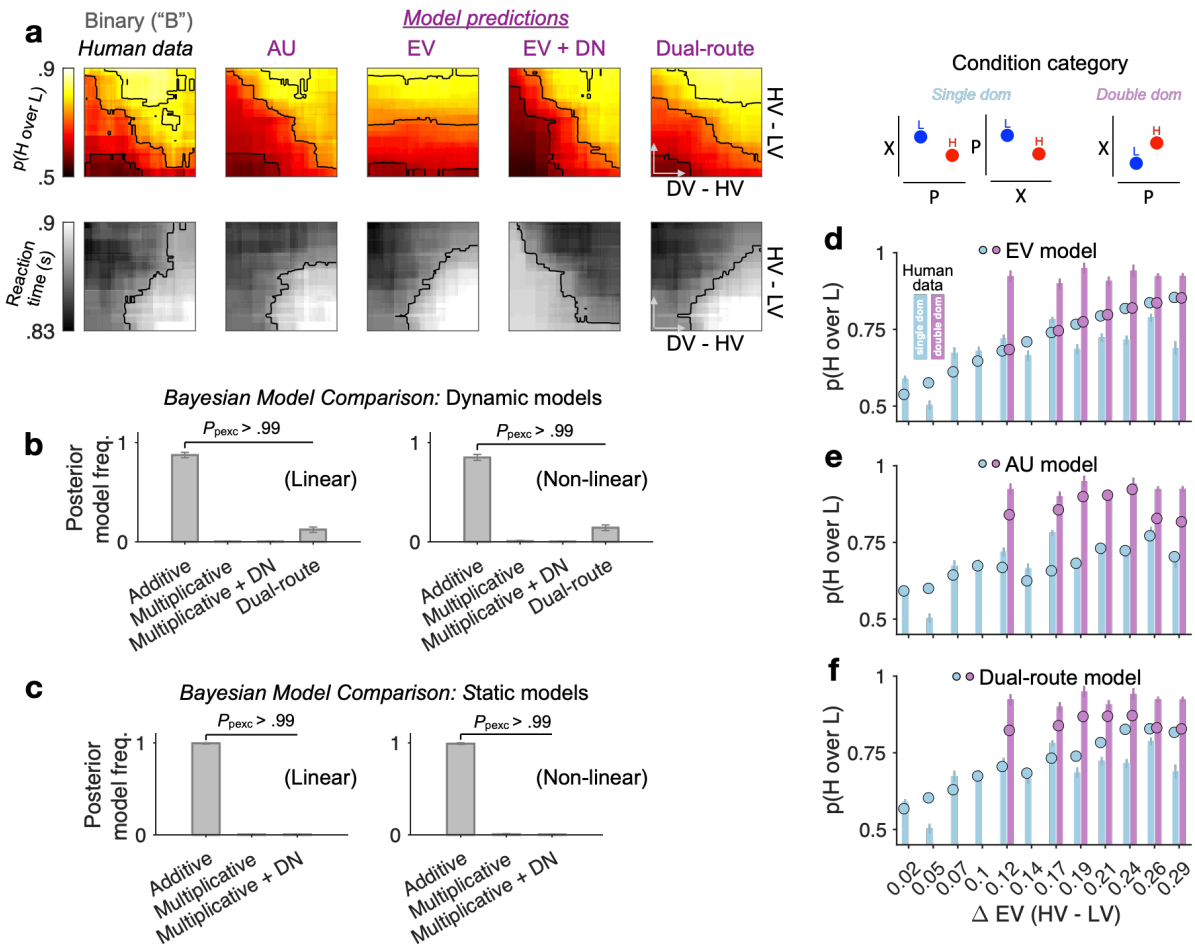


Figure 3. Modelling binary choice behaviour. (**a**) Human vs. model-predicted choice accuracy and RT patterns. (**b**, **c**) Bayesian model comparison based on cross-validated log-likelihood. Linear and non-linear: different psychometric transduction functions for reward magnitude and probability (see Methods). Linear multiplicative: EV; Linear additive: AU. Dynamic model: optimisation based on a joint likelihood of choice probability and RT; Static model: optimisation based on binomial likelihood of choice probability. (**d**–**f**) Model predictions (circles) plotted against human data (bars) for different condition categories (H dominates L on one attribute only, i.e., *single dom*. (81% of all trials), or on both attributes, i.e., *double dom*. (only 19%)) at each level of HV – LV. Error bars = ± s.e.m. (N = 144 participants).

Next, we used formal Bayesian model comparison to quantitatively identify the simplest possible model that can reproduce the binary-choice behaviour. For completeness, we added in the comparison a naïve expected-value (EV) model, and a recently proposed dual-route model, which relies on EV utility and flexibly arbitrates between DN and mutual-inhibition processes[18]. A vanilla DN model can be viewed as a nested model within the dual-route model. All models were fitted to each participant's binary choices and reaction times (RTs) by optimising a joint likelihood function (Methods). Qualitatively, **Fig. 3a** shows that the naïve EV model fundamentally deviates from the human choice data: The model predicts a vertical gradient of choice accuracy constrained by HV – LV. When comparing these models head-to-head using a cross-validation scheme, we find that the AU model wins over any other model (**Fig. 3b**; protected exceedance probability P_{pexc} > .99; Methods). This result still holds robustly when including, in all models, subjective non-linear distortions of attributes[34] ('Non-linear' in **Fig. 3b**, **c** right panels; Methods). Moreover, the RTs predicted by the divisive normalisation (DN) model mismatch the human RTs (**Fig. 3a**: 'EV + DN'). But this is not the sole reason why

DN fails in the model comparisons. DN still fails to win over the AU model even when we consider static versions of the models fitted to the choice probabilities alone while ignoring the RTs (**Fig. 3c** 'Static models'; Methods). These systematic model comparisons thus quantitatively support the AU model as a remarkably simple account of the 'notional distractor' effects in binary trials (**Fig. 2i**). Additional model comparisons corroborate that the AU model is favoured over specific and popular instantiations of the non-linear multiplicative model (i.e., expected utility[20] and prospect theory[35]; see **Figure 3—figure supplement 1**).

The success of the AU model suggests that the reward attributes were not multiplicatively conglomerated into a single EV dimension. To better illustrate the differences between AU- and EV-based models, we plotted model predictions separately for different condition categories in which H dominates L in one attribute only (single dom.) or in both attributes simultaneously (double dom.; only 19% of all trials) (**Fig. 3d—f**). The EV model, being only sensitive to the difference in the product of probability and reward, fails to capture any performance variations across these distinct categories (because these categories are orthogonal to the EV difference; **Fig. 3d**). By contrast, the AU model can capture these across-category performance changes (**Fig. 3e**). Qualitatively, the dual-route model seems to be able to capture some patterns of the human data on the binned maps (**Fig. 3a**: 'Dual-route'), but its deviation from human behaviour becomes more salient when inspecting the model predictions at each fixed level of EV difference (**Fig. 3f**). Taken together, we show that the AU model reproduces the patterns of both accuracy and RT in binary-choice trials decisively better than the models relying on EV or divisively normalised EV (also see **Supplemental Tables 1** and **2** for model parameter estimates.

A modest negative distractor effect operating on additive utility

Because 'additive integration' is a decisively better model of binary-choice behaviour than multiplicative models are (e.g., EV, EV + DN, expected utility, and prospect theory), here we re-assessed distractor effects after defining the key analysis variables using the subjective utilities afforded by the AU model. To better define these variables, we fit the dynamic AU model (with free λ weight; Methods: *Context-independent model*) to binary and ternary choices, separately, and constructed subjective utility as AU = λ(X) + (1 – λ)(P). We then analysed relative choice accuracy using a GLM with AU-based regressors: $AU_H - AU_L$ (AU difference between targets), $AU_H + AU_L$ (AU sum), distractor variable (absolute: AU_D, or relative: $AU_D - AU_H$), distractor variable x ($AU_H - AU_L$) interaction, and the interaction between each of these effects and 'C' (Condition: T vs. B). We focus on the key regressors that assess the distractor effects in ternary choices over and above the matched baseline performance (captured by the 'x C' interaction): the main distractor effect, 'distractor variable x C', and the interacting distractor effect dependent on choice difficulty, 'distractor variable x ($AU_H - AU_L$) x C'. First, we verified that including the AU-sum regressors decisively improved the GLM's goodness-of-it, $\chi^2(2) > 1.005 \times 10^3$, $p < .0001$; and the GLM with AU regressors had better goodness-of-it than the GLM with EV regressors, $\chi^2(0) > 1.03 \times 10^4$, $p < .0001$ (for either absolute or relative distractor variable). We found a significant *negative* main distractor effect, AU_D x C, on the choice accuracy, $t(143) = -2.27$, $p = .024$ (uncorrected); however, this main effect was marginally significant when using the relative distractor variable $AU_D - AU_H$, $t(143) = -1.86$, $p = .065$ (uncorrected). The interacting distractor effect was not significant $t(143) < -.54$, $p > .1$ (see **Supplemental Table 3** for a full report). We here reiterate that, by contrast, neither the main nor the interacting distractor

effects were significant in the GLM with EV-based regressors ($p > .3$, uncorrected, **Fig. 2f** and **Figure 2—figure supplement 2**, **3**). These results based on subjective AU fit more closely with recent reports showing a *negative* main distractor effect[5].

Distractor-related covariations and multiattribute context effects

Having revisited distractor effect while controlling for target-related confounding covariations, we now move into examining distractor-related covariations, i.e., whether multiattribute context effects were present in these datasets, and if so, whether they were related to the modest AU-based negative distractor effect we reported above.

Classic multiattribute context effects typically occur when a decoy alternative is added to a choice-set consisting of the two non-decoy alternatives (**Fig. 4a**). Two classic context effects are particularly relevant in this task because 84% of the ternary trials satisfy specific H, L, and D geometric (i.e., in the two-dimensional choice space) configurations that could give rise to these context effects (**Fig. 4a**). First, an attraction effect[3], whereby the probability of choosing a target over a competitor increases due to the presence of an 'inferior' *dominated* D near the target (also see **Fig. 1b—d** for an illustrative example); and second, a repulsion effect[6], whereby this pattern is reversed (and the probability of choosing the target is reduced) in the presence of a nearby 'superior' *dominating* D. We note in passing that other popular context effects, such as the similarity and the compromise effects, could not be examined here due to the absence or the very low frequency of occurrence of appropriate choice-sets (i.e., in these effects the decoys are required to be roughly iso-preferrable to the target or the competitor alternatives[36]).

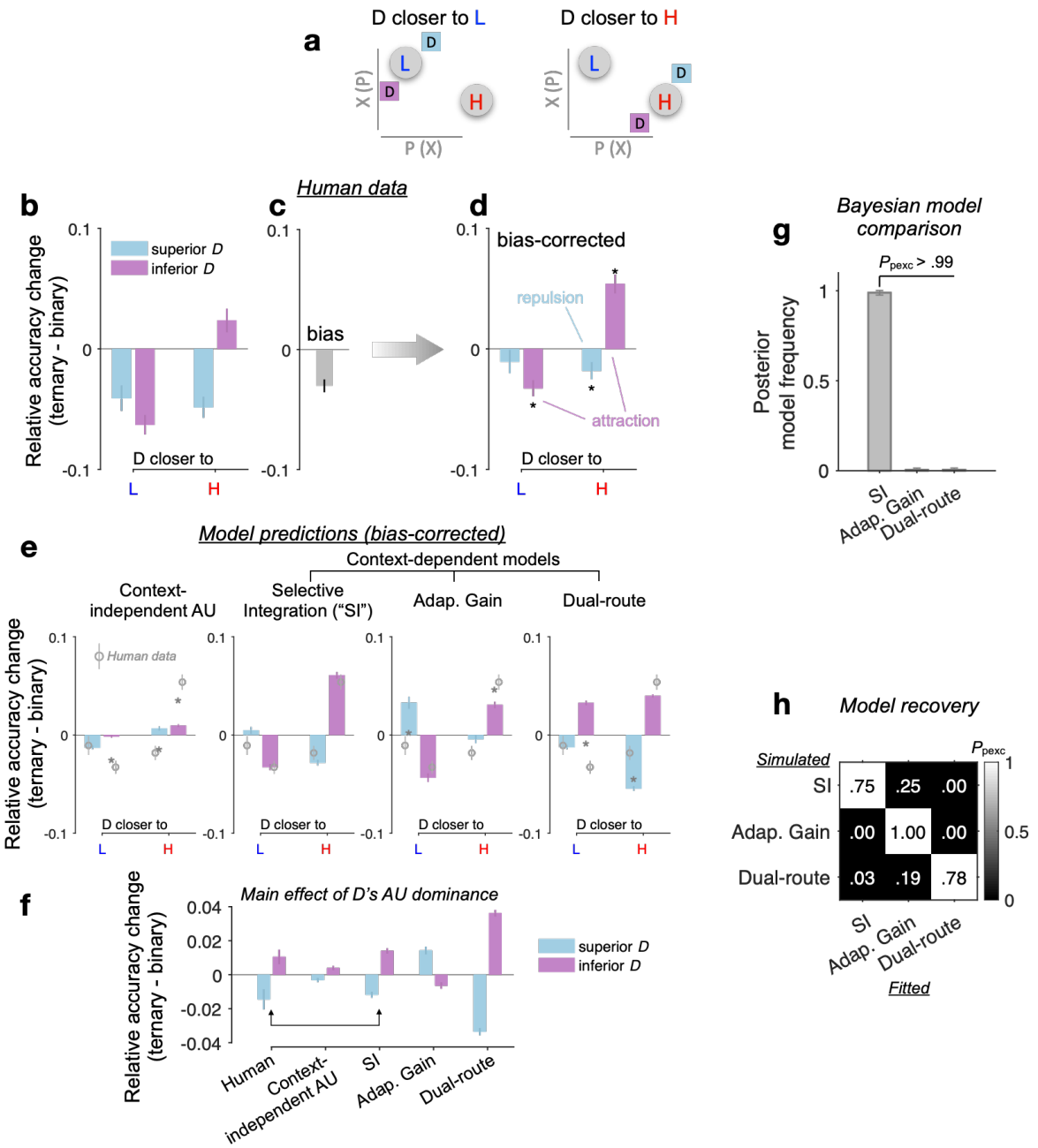


Figure 4. Multiattribute context effects. (**a**) The attraction effect[3] can happen when D is close to and dominated by H or L. The repulsion effect[6] can happen when D is close to but dominates H or L. (**b**) T – B *p*(H over L) plotted as a function of D's relative Euclidean distance to H vs. L in the attribute space and D's AU dominance – 'superior' ('inferior') D lies on a higher (lower) AU indifference line than both H and L. Error bars: ± s.e.m. across participants (*N* = 144). (**c**) Condition-unspecific bias of *p*(H over L) in ternary conditions relative to binary conditions. (**d**) Bias-corrected T – B *p*(H over L). Black asterisks: significant context effects (Bonferroni-Holm corrected $p < .05$; two-tailed 1-sample *t*-tests against 0). (**e**) Bias-corrected T – B *p*(H over L) predicted by different models. Grey asterisks: significant difference between model predictions and human data (uncorrected $p < .05$; two-tailed paired-sample *t*-tests). (**f**) Main effect of D's AU dominance (superior vs. inferior) on T – B relative accuracy change in human data and model predictions. (**g**) Bayesian model comparison of context-dependent models (also see Figure 4—figure supplement 1). (**h**) Model recovery (Methods). Text in each cell: posterior model frequency; Heatmap: protected exceedance probability (P_{pexc}) that a fitted model (each column) has the highest log-likelihood given input ternary-trial data (choice probabilities and RTs) simulated by a certain model (each row).

Before characterising decoy effects, we considered the possibility that the change of relative accuracy between binary (B) and ternary (T) choice reflects a condition-unspecific change in behaviour, i.e., an accuracy reduction potentially induced by higher cognitive demands in the ternary trials. We estimated this 'generic' bias by a permutation procedure whereby we randomly shuffled the true mappings between ternary conditions and their matched binary conditions and obtained an average T – B accuracy contrast across random permutations (**Fig. 4c**; see Methods). After this bias was corrected for each participant, the relative accuracy change from B to T was compared against zero (i.e., a null hypothesis: T and B had no difference over and above the condition-unspecific bias). The bias-corrected T – B relative accuracy change shows a strong attraction effect: The relative choice accuracy dynamically decreases and increases as an inferior D gets closer to L (n = 32 relevant trials in this category) and H (n = 33), respectively, $|t(143)| > 4.88$, $p < .0001$ (**Fig. 4d**). Meanwhile, a significant repulsion effect occurs when a superior D dominates H or L in the attribute space, $t(143) = -2.53$, $p = .025$, particularly when D is closer to H (n = 41 relevant trials in this category), but not when D is closer L (fewer trials n = 20 in this category; $t(143) = -1.09$, $p = .28$, Bonferroni-Holm corrected). A repeated-measures ANOVA reveals a significant interaction between the similarity (D closer to H vs. D closer to L) and the dominance of D (inferior vs. superior), $F(1, 143) = 30.6$, $p < .0001$, partial $\eta^2 = .18$, on the T – B relative accuracy change.

Of note, in line with the AU-based negative distractor effect we reported in the previous section, the current analysis on multiattribute decoy effects also reveals a significant main effect of D's AU dominance on relative accuracy change (repeated-measures ANOVA on T – B relative accuracy change in all trials), $F(1, 143) = 6.51$, $p = .012$, partial $\eta^2 = .044$ (**Fig. 4f**: 'Human'). This overall dominance effect seems to be induced by a noticeable asymmetry in the observed decoy effects (**Fig. 4d**, net of inferior-purple vs. superior-blue bars): targeting H seems more effective than targeting L, and also, the attraction effect is stronger than the repulsion effect (hence, the multiattribute effects do not cancel out but lead to a net accuracy boost for inferior D's). Crucially, across participants there is a significant positive correlation between the negative main distractor effect 'AU_D x C' (GLM in the previous section) and the negative main effect of D's AU dominance here (Spearman's $\rho = .28$, $p = .00074$). This indicates that if a participant had a stronger *asymmetry* in the decoy effects (D's AU dominance) then they would also have a stronger AU-based negative distractor effect. We further interpret this correspondence using computational modelling in the next section.

Selective gating explains multiattribute context effects and adapts to decision noise

Although the AU model successfully described behavioural responses in binary trials (**Fig. 3**), it is essentially a context-independent model complying with the *independence from irrelevant alternatives* (IIA) principle. That is, according to the AU model, the characteristics of D in ternary trials should not lead to any preference change between the two available targets H and L. In this section, we resort to computational modelling to understand why multiattribute context effects occur in ternary trials. First, as shown in **Fig. 4e** leftmost panel, the T – B effects in human data could not be explained by a context-independent AU model ('blind' to D, i.e., only taking H and L as inputs) with freely varying parameters across B and T conditions (the free parameters could capture any global changes in behaviour due to, e.g., different noise

levels and/or attribute weighting, potentially induced by higher cognitive demands in the ternary trials; Methods).

Next, we considered two models that can predict context effects in multiattribute choice: selective integration (SI) and adaptive gain[6,37] (Methods). In the SI model, the gain of processing on each attribute value relies on its rank within that attribute, with a selective gating parameter *w* controlling the suppression strength for the low-rank attribute value. In the adaptive gain model, each alternative is evaluated relative to the context statistics via a non-linear function. We compared these models to the context-dependent dual-route model by Chau et al., which had not been previously examined in relation to classic context effects[18]. As shown in **Fig. 4e**, the SI model can reproduce all types of attraction and repulsion effects in the data, but a dual-route model cannot. A systematic Bayesian random-effects model comparison shows that SI prevails over all other models ($P_{pexc} > .99$; **Fig. 4g**), including the dual-route, DN, or any context-independent model with either EV or AU as the utility function (**Figure 4—figure supplement 1a**). The adaptive gain model captures some key context effects, although not as closely as the SI model does (fixed-effects: SI vs. adaptive gain, cross-validated log-likelihood ratio = 219.8, ΔBIC = -1194, pooled over participants; random-effects: SI's $P_{pexc} > .99$; **Fig. 4e**), and explains the human data decisively better than the dual-route model (adaptive gain vs. dual-route: cross-validated log-likelihood ratio = 48.58, ΔBIC = -806, pooled over participants; adaptive gain's $P_{pexc} = .78$; see **Supplemental Table 4** for model parameters). A model recovery procedure (Methods) also confirmed that our Bayesian model comparison was capable of properly discriminating among these context-dependent models (SI, adaptive gain, and dual-route; **Fig. 4h**).

Crucially, the SI model could lawfully reproduce the overall negative D's AU dominance effect without relying on any additional mechanism such as divisive normalisation (DN), $F(1, 143) = 78.4$, $p < .00001$, partial $\eta^2 = .35$, post-hoc *t*-test: $t(143) = -8.85$, $p < .0001$ (**Fig. 4f**). By contrast, the models that predict 'by-design' unidimensional distractor effect (dual-route and DN) could not describe the multiattribute decoy effects (**Fig. 4e**, **f**; also see **Figure 4—figure supplement 1b**, **c**). This suggests that the negative D dominance effect is not the by-product of a unidimensional distractor mechanism but instead falls out naturally from the presence of asymmetrical decoy effects (i.e., stronger attraction effect, when H is targeted). The correlation between the D dominance and AU distractor effects discussed above, corroborates the possibility that, to some extent, the negative distractor effect is driven by asymmetrical decoy effects (**Fig. 4f**: SI model mimics the negative D dominance effect).

How does the winning selective integration (SI) model capture the attraction and repulsion effects? The way the model works in these two effects is shown in a schematic form in **Fig. 5a** and **b**. Because SI selectively allocates reduced gain to 'locally losing' attribute values, it exerts a stronger gating on the utility of option B when the decoy D dominates B (**Fig. 5b**) than when D is dominated by B (**Fig. 5a**). Thus, the relative preference between A and B can be dynamically reversed according to the dominance of D, giving rise to the opposing context effects of attraction and repulsion.

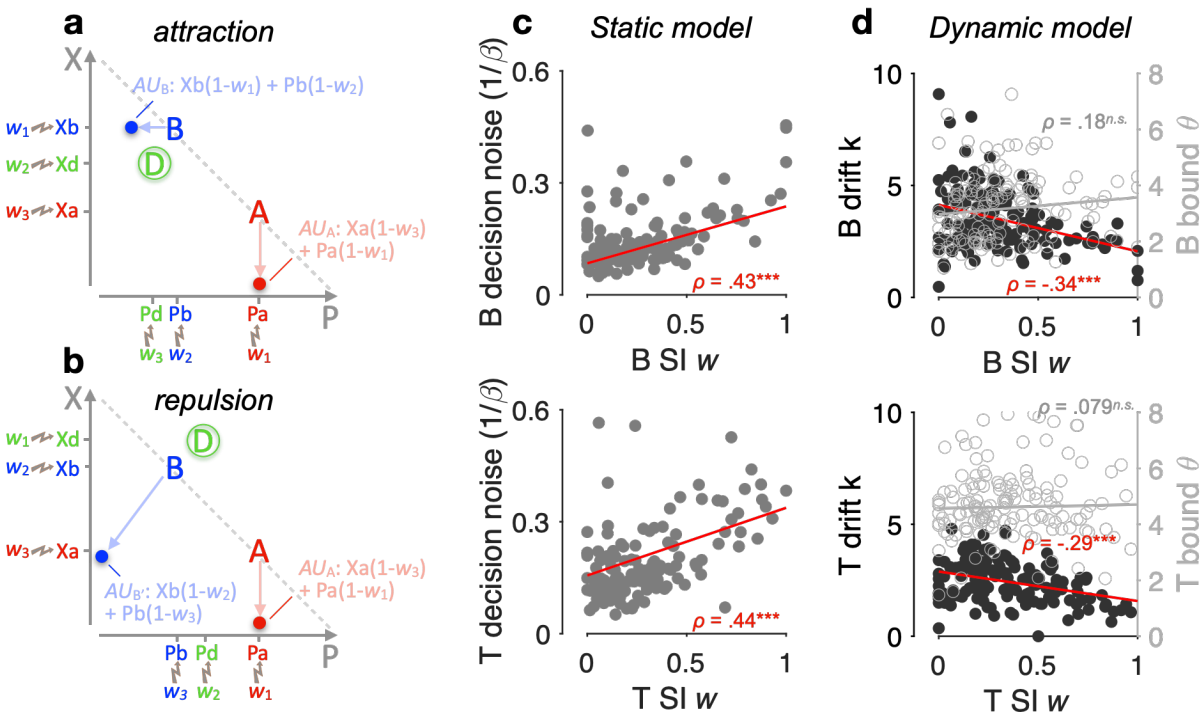


Figure 5. Selective gating, context effects, and decision noise. (**a**, **b**) By virtue of SI, alternatives A, B, and D are ranked within each attribute dimension; SI weights are then assigned to the sorted alternatives accordingly (for illustration, $w_1 = 0$, $w_2 = .5$, and $w_3 = .9$). Attribute values are selectively suppressed in proportion to the SI weights, i.e., lower ranked values are suppressed more strongly. As a result of SI (assuming any $w_1 < w_2 <= w_3$), the additive utility (AU) of B is higher in the case of panel a than that of B' in the case of panel b due to the distinct locations of decoy D. (**c**) Relationship between SI w and decision noise. Red lines: robust linear regression fits (using Matlab's robustfit). β: softmax inverse temperature. Top: binary trials ('B'); bottom: ternary trials ('T') (ternary SI w: mean across w_2 and w_3); SI $w = 0$: lossless integration, i.e., attributes maintain their original values; SI $w > 0$: selective integration; SI $w = 1$: attribute value shrinks to 0. ρ: Spearman's rank correlation coefficient. $^{***}p < .0001$. (**d**) Relationship between SI w and drift sensitivity k (filled black) or decision bound (open grey).

Lastly, selective integration has been shown to maximise decision accuracy and economic outcomes in the face of decision noise that arises when multiple attributes or pieces of information need to be combined towards a decision[31,38]. Here, we examined whether SI confers robustness on decisions by adapting to the individual decision noise levels. And if so, it could provide a normative justification of these decoy biases as emergent properties of a decision policy that maximises robustness against decision noise. After fitting a static version of the SI model to choice probabilities (Methods), we found a positive correlation across participants between the SI gating w and the temperature of a softmax choice function (Spearman's $\rho(142) > .43$, $p < .0001$, for both binary and ternary choices; **Fig. 5c**). Additional parameter-recovery simulations corroborated that these findings are not due to any trade-off between model parameters (**Figure 5—figure supplement 1**). After fitting the dynamic SI model to choices and RTs together, we found that the SI w negatively correlates with the utility drift sensitivity (how effective a participant is in using the information about utility; Methods), but not the height of the decision bound (**Fig. 5d**). These findings suggest that SI has an adaptive role, as it selectively exaggerates the attribute difference between winning and losing alternatives by just the right amount in proportion to decision noise.

Discussion

Context effects violate the axioms of rational choice theory and can shape our understanding of the computational mechanisms that underlie decision-making. Recent reports suggested that such effects do not occur only in multiattribute space, but also at the level of a single value dimension[4,5,18,39]. However, these reports diverge in their conclusions[16,19,40], with some studies showing a negative distractor effect while others, using risky-choice tasks, showing positive and interacting distractor effects. Here we focused on these risky-choice datasets and examined whether the positive and interacting distractor effects could afford alternative interpretations. We found that the previously reported positive and interacting effects can be explained away by confounding covariations among distractor variables and subjective utility, which in these datasets was shaped by additive and not multiplicative reward/probability integration. When we redefined distractor effects using the descriptively more adequate 'additive utility', we instead reported a modest negative distractor effect. Asides from this modest negative distractor effect, classic context (decoy) effects—determined by the multiattribute similarity and dominance relationship between the distractor and the available alternatives—occurred robustly. Of note, the multiattribute context effects occurring here were technically 'phantom' effects[41] due to the fact that the distractor could not be chosen; and contrary to most previous demonstrations, where the target and competitor alternatives are almost equally valuable, the effects we reported here arise even under a noticeable value difference between the target and the competitor[42]. These context effects were asymmetric in magnitude, yielding an average accuracy increase for inferior distractors and an average accuracy decrease for superior ones, mirroring thus the negative distractor effect.

These results clarify the role that an unavailable distractor alternative plays in decisions between two uncertain prospects. Notably, the negative distractor effect we reported is of rather small size. A plausible explanation of this negative distractor effect is that it could be driven, to some extent, by asymmetric multiattribute context effects. We made this conjecture for two reasons. First, from an empirical perspective, the 'dominance' effect arising from multiattribute decoy effects correlates with the negative distractor effect. Second, from a theoretical perspective, model comparison favours by large a model that can only produce multiattribute decoy effects and has no dedicated mechanism to generate unidimensional distractor effects. This alludes to a parsimonious account that precludes genuine unidimensional distractor effects. However, we interpret the link between the distractor and decoy effects with caution given that there are differences in the way the two effects were quantified. In the distractor effect, the key variable is defined continuously (i.e., the distractor's utility) while in decoy effects, 'dominance' is defined categorically (superior or inferior to the target). Similarly, a less parsimonious hybrid model in which both unidimensional distractor mechanism (i.e., divisive normalisation) and multiattribute context-sensitivities (i.e., selective integration) coexist has not been ruled out. Definitive conclusions on the link between distractor and decoy effects await new experimental designs in which the two types of effects can be quantified independently.

In light of the stronger robust presence of decoy effects in these datasets, the hypothesis that multiattribute complexity is a key moderator of context dependencies, remains viable[11,43,44]. Encoding absolute information along several attributes is inefficient[2] and suffers from the *curse of dimensionality*[45,46]. A plausible solution to circumvent these issues is to discard information

592 that is less important[47]. In a similar spirit, our selective integration (SI) model preferentially
593 focuses on larger values at the expense of smaller values, and because of this policy, context-
594 effects ensue. This selective policy potentially maximises robustness to decision noise[31,48],
595 providing thus a normative justification of contextual preference reversal. And although the
596 selective integration model had thus far been applied to dynamic (sequential) value
597 accumulation tasks[49], we here show that it can also account for behaviour in a static
598 multiattribute task. Nevertheless, we do not claim that SI, despite its simplicity and
599 computational tractability, is the only computational model that could explain the
600 multiattribute context effects in this task[50–53].
601
602 Notably, and contrary to the normative rule, participants in the datasets we analysed
603 integrated probability and payoff information additively and not multiplicatively. This additive
604 integration, or more generally, the processing of information in a 'within dimensions' fashion,
605 is probably what spurs contextual preference reversal in risky choice. However, unlike what we
606 observed here, a large and influential body of research in risky decision-making supports the
607 idea that non-linear transformations of probabilities and rewards are multiplicatively
608 combined (e.g., prospect theory[35]). Additive integration in risky choice has limited applicability
609 as opposed to the more widely applied multiplicative-integration framework. Unlike
610 multiplicative models, the additive-integration framework enforces independent contributions
611 of reward and probability[54] and cannot readily accommodate rich attitudes towards risk[55] or
612 other complex phenomena in decisions over multibranch gambles[56,57]. At an even more
613 fundamental level, additive integration requires ad-hoc assumptions (on the way outcomes
614 and probabilities are represented) to capture the decreasing desirability of negative outcomes
615 with increasing probability, or risk-attitudes over mixed gambles that offer both positive and
616 negative outcomes[58]. Although these limitations suggest that additive integration is not readily
617 generalisable to all types of risky choice, this does not undermine its descriptive superiority in
618 simpler tasks like the one we focused on here. One overarching conjecture is that the stimulus
619 employed in different experiments determines the way participants integrate reward and
620 probability, with multiplicative integration observed in numerically explicit gambles and
621 additive integration when rewards and probabilities are inferred from abstract non-numerical
622 (perceptual) dimensions[25–27,30,54]. This may further suggest that risk preference is a rather
623 elusive and unstable construct[59].
624
625 Additive and multiplicative integration strategies are not necessarily mutually exclusive, and
626 they could instead co-influence risky decisions[27,30]. More generally, neural correlates of both
627 additive and multiplicative decision variables have been reported within the same task and
628 even within overlapping brain regions. Neural signals in a network of cortical and subcortical
629 brain regions have been shown to code for value/utility-related metrics of diverse assets[58,60–
630 62]. Within this network, the dorsal medial frontal cortex (dmPFC in humans[11] and an
631 anterior/dorsal frontal site located between the rostral and anterior cingulate sulci in area 32
632 of non-human primates[30]) seems to be particularly relevant for the multiplicative integration
633 of reward and probability in risky decisions. Strikingly, a focal ultrasonic disruption of this
634 anterior/dorsal frontal region (but not the posterior/ventral medial frontal cortex) caused a
635 shift from a multiplicative to an additive strategy in the macaques[30], imputing a direct causal
636 link between EV computation and the dorsal medial frontal lobe. By contrast, the ventral
637 medial frontal cortex (vmPFC) seems to be involved in representing individual sources of
638 reward information such as the latent reward probabilities or states[25,63] and the prospective

appetitive values of risky gambles[25], and is causally linked to the within-attribute comparison across alternatives[44], providing key process components for the additive strategy. Interestingly, however, the human dmPFC, but not vmPFC, has been shown to guide risky choice by combining *additively* bandits' reward probabilities with their normalised utilities[25]. Explicit coding of multiplicative value has also been found in the orbitofrontal cortex (OFC) and ventral striatum[64], while reward attribute representations and comparisons could also engage the OFC[65] or the parietal cortex[11] depending on the task and the relevant attributes. This rather complicated picture of the neural coding of multiplicative vs. additive integration suggests that, even though behaviour may be by and large best described by a particular decision strategy[66], the brains could afford simultaneous representations of multiple competing strategies[66–68].

To conclude, harnessing the power of open science, we scrutinised large datasets from published studies (over 140 human participants) and re-assessed the role of the distractor alternative in decisions under uncertainty. More broadly, we adopted an integrative approach attempting to characterise the interplay between two types of context effects (unidimensional and multiattribute), which had thus far been examined in isolation. Our findings contribute to a large body of research[32,50,69,70] that challenges the standard tenet of neuroeconomics that value representation is menu-invariant[71–73].

Methods

Multiattribute decision-making task and datasets

In the multiattribute decision-making task of Chau et al.[4], participants made a speeded choice within 1.5 s between two available options (HV: high value; LV: low value) in the presence ('ternary') or absence ('binary') of a 3rd unavailable distractor option (D). On each trial, the 2 or 3 options appeared in different quadrants of a computer screen. Each option was a small rectangle and its colour/orientation represented some level of reward magnitude/probability (the feature-to-attribute mapping was counterbalanced across participants). There were equal number of ternary (n = 150) and binary trials (n = 150). However, the ternary and binary trials did not have 'one-to-one' mapping, i.e., some ternary trials had only 1 matched binary trial with the same H and L options while others had several (also see **Figure 2—figure supplement 1**). The available targets and the unavailable distractor were flagged by orange and pink surround boxes, respectively, 0.1 s post stimulus onset. Participants received feedback about whether the chosen option yielded a reward at the end of each trial. We re-analysed 5 published datasets of this speeded-decision task (*N* = 144 human participants in total). These 5 datasets all include ternary and binary trials: the original fMRI dataset (*N* = 21[4]), a *direct* replication study (using exact trial sequence provided by Chau et al.) performed by Gluth et al. (Gluth Experiment 4: *N* = 44), and additional replication experiments performed by Gluth et al. (Gluth Experiments 1, 2, and 3; *N* = 79 in total[19]). Of note, a new dataset ('Hong Kong'; *N* = 40) recently acquired by Chau et al.[18] did not contain binary trials and thus could not be included in our analyses that compared ternary and binary choices head-to-head.

Computational modelling of behaviour

General architecture of dynamic models

We used the framework of feedforward inhibition (FFI) model to jointly fit choice probability and RT. The FFI model conceptually 'interpolates' between a race model and a drift-diffusion model[74]. Each accumulator is excited by the evidence for one alternative and inhibited by the evidence for the other alternatives via feed-forward inhibition. Evidence units are ideal integrators with no leak, and they receive crossed inhibition from different alternatives. The activation of each integrator is x_i (for the ith alternative), which takes I_i as evidence input:

$$dx_i = (kI_i + I_0)dt + \xi\sqrt{dt} \tag{1}$$

Where k is a sensitivity parameter (drift rate coefficient, i.e., how effective the nervous system is in using utility information), and I_0 a baseline drift, independent of I_i. ξ denotes Gaussian noise of a zero mean and a variance of σ^2. The noise term is in proportion to the square-root of dt because the noise variance itself is a linear function of dt. Evidence I_i is computed as the utility of alternative i inhibited by the average utility across all other alternatives in a choice set (assuming *n* alternatives in total including i):

$$I_i = U_i - c\left(\frac{\sum_{j\neq i} U_j}{n-1}\right) \tag{2}$$

Where c ($0 \leq c \leq 1$) is the FFI strength. When c = 0, the model becomes a classic race model; when c = 1, it corresponds to a drift-diffusion model. The drift rate μ of FFI diffusion thus can be rewritten as:

$$\mu_i = kU_i - ck\left(\frac{\sum_{j \neq i} U_j}{n-1}\right) + I_0 \tag{3}$$

The utility function U_i can take different forms of interest, such as expected value (EV), additive utility (AU), or being adaptive to contexts (SI and adaptive gain models, see next).

We next derive the analytical solutions for the FFI dynamic system (equation 1). First, the distribution of first-passage times t (i.e., decision-bound hitting times) is known to follow an inverse Gaussian distribution[75]:

$$p(t; \theta, x_0, \mu) = \frac{\theta - x_0}{\sqrt{2\pi\sigma^2 t^3}} e^{\frac{-(\theta - x_0 - \mu t)^2}{2\sigma^2 t}} \tag{4}$$

In which, x_0 is a starting point of the integrator, and θ, the height of decision bound. Then, we can derive the probability of observing a response time RT on a given trial as:

$$p(t = T) = \frac{\theta - x_0}{\sqrt{2\pi\sigma^2 T^3}} e^{\frac{-(\theta - x_0 - \mu T)^2}{2\sigma^2 T}} \tag{5}$$

Where $T = RT - t_{\mathrm{nd}}$, and t_{nd} is a non-decision response delay. Now, we can obtain the probability of observing a choice i but not choice j at time T as:

$$p(\text{choice} = i, t = T) = p(t_i = T)p\big(t_j > T\big) = p(t_i = T)\big[1 - p\big(t_j < T\big)\big] \tag{6}$$

That is, we can multiply the PDF for response i (e.g., H choice) with the cumulative probability that the alternative response j (e.g., L choice) hits the decision bound if and only if at a time later than *T*. Given that a CDF of an inverse Gaussian is expressible in terms of the CDF of the standard Gaussian $\Phi(\cdot)$, i.e.,

$$p\big(t_j < T\big) = \Phi\left(\frac{\mu_j T - \theta + x_0}{\sigma\sqrt{T}}\right) + e^{\frac{2(\theta - x_0)\mu_j}{\sigma^2}} \Phi\left(\frac{-\mu_j T - \theta + x_0}{\sigma\sqrt{T}}\right) \tag{7}$$

and by substituting equation (7) into equation (6), we obtain the likelihood function of observing choice i at time T as:

$$L(i, T) = \left[1 - \Phi\left(\frac{\mu_j T - \theta}{\sigma\sqrt{T}}\right) + e^{\frac{2\theta\mu_j}{\sigma^2}} \Phi\left(\frac{-\mu_j T - \theta}{\sigma\sqrt{T}}\right)\right] \frac{\theta}{\sigma\sqrt{2\pi T^3}} e^{\frac{-(\theta - \mu_i T)^2}{2\sigma^2 T}} \tag{8}$$

Further, we assume no bias in the starting point of accumulation, i.e., $x_0 = 0$, and a unit variance of the noise term, i.e., $\sigma = 1$. This is equivalent to measuring the diffusion properties such as sensitivity k and bound θ in units of σ, a convention in standard diffusion models that prevents parameter trade-offs via a re-scaling of the diffusion space[76,77]. When fitting the models to data, we maximised the sum of Log(*L*) (*LL*, hereafter) across trials given empirical trial-by-trial choices and RTs. From this, and after getting the best-fitting model parameters,

we can get the model prediction of choice probability (equation 9) and mean RT (equation 10) using the following integrals, respectively:

$$p(\text{choice} = i) = \int_0^{+\infty} L(i,t)dt \tag{9}$$

$$\mathbf{E}(t_i) = \int_0^{+\infty} \big(tL(i,t)\big)dt \Big/ \int_0^{+\infty} L(i,t)dt \tag{10}$$

In practice, we used the trapezoidal numerical integration with a small enough dt = .001 s and large enough T_{max} = 100 s. Empirically no negative drift rates were observed in any of the model parameter estimates in any participant.

Context-independent model

The utility function U_i in the context-independent model is 'blind' to the distractor (D). When being fitted to ternary trials, the context-independent model only takes H and L as inputs. For instance, the context-independent AU model assumes additive utility, AU_H = (λ)HX + (1 - λ)HP, and AU_L = (λ)LX + (1 - λ)LP, with λ being the attribute weight of reward magnitude (0 ≤ λ ≤ 1).

Selective integration (SI) model

We adopted the simplest form of SI model. The SI model discards information about choice alternatives with relatively lower value in each attribute separately. Alternatives are compared and then selectively gated based on their ranks within each attribute. The highest attribute value remains unchanged ($w_1 = 0$), intermediate value gets suppressed by a multiplicative weight w_2, whilst the lowest value gets more strongly suppressed by w_3 ($0 \leq w_2 \leq w_3 \leq 1$). That is, the SI w represents the percent reduction in information processing gain in an attribute. In case of a tie for the highest value, the two high values are unchanged, and the lowest value is suppressed by w_3. By contrast, in case of a tie for the lowest value, the two low values both get suppressed by w_3. After the selective gating within each attribute, which transforms attribute values to X' and P', respectively, an additive utility is constructed for each alternative via a weighted averaging, i.e., AU = (λ)X' + (1 - λ)P', with λ being the attribute weight of reward magnitude (0 ≤ λ ≤ 1). This additive utility (after SI) is then fed into the FFI model as the component U for each alternative (equation 3).

Adaptive gain model

The attribute value of a certain alternative, i, e.g., the reward magnitude, X_i, is inhibited by a context mean across all alternatives in a choice set (e.g., *n* alternatives), which yields $\overline{X_i}$. $\overline{X_i}$ then passes through a sigmoidal transduction function with slope *s* and bias term *b*. This gain control is adaptive in that it accommodates the contextual information. Importantly, the gain control operates independently on each attribute, and it produces the final subjective attribute value (X_i^{AG}) for the construction of additive utility for each alternative:

$$X_i^{AG} = 1 \Big/ \left(1 + e^{\frac{b_x - \overline{X_i}}{s_x}}\right) \tag{11}$$

$$\overline{X_i} = X_i - \left(\frac{\sum_i X_i}{n}\right) \tag{12}$$

The adaptive gain modulation of P can be obtained by replacing X with P in the above two equations. Ultimately, the utility of each alternative is computed as a weighted sum across attribute values:

$$U_i = \lambda X_i^{AG} + (1-\lambda) P_i^{AG} \tag{13}$$

Dual-route model

The dual-route model arbitrates between a vanilla mutual inhibition (MI) process and a MI process with additional divisive normalisation (DN). In the vanilla MI 'route', this model assumes that the evidence input to the accumulator is:

$$I_i = U_i - f_{MI}\left(\frac{\sum_i U_i}{n}\right) \tag{14}$$

Where U is expected value (EV), i.e., EV = XP, and f_{MI} is the inhibition strength of MI. In another rival route, the utility of each choice alternative is additionally divisively normalised by the sum of utilities across all alternatives (H + L in a binary trial; H + L + D in a ternary trial):

$$I_i^{DN} = U_i^{DN} - f_{MI}\left(\frac{\sum_i U_i^{DN}}{n}\right) \tag{15}$$

$$U_i^{DN} = U_i \Big/ \sum_i U_i \tag{16}$$

The drift rate takes the same form as that in FFI: $\mu_i = kI_i + I_0$. Based on equation (8) and for clarity, we rewrite the two components in (8) as $\mathbf{g}$ and $\mathbf{G}$:

$$\mathbf{G}(\mu, T, \theta, \sigma) = 1 - \Phi\left(\frac{\mu T - \theta}{\sigma\sqrt{T}}\right) + e^{\frac{2\theta\mu}{\sigma^2}} \Phi\left(\frac{-\mu T - \theta}{\sigma\sqrt{T}}\right) \tag{17}$$

$$\mathbf{g}(\mu, T, \theta, \sigma) = \frac{\theta}{\sigma\sqrt{2\pi T^3}} e^{\frac{-(\theta-\mu T)^2}{2\sigma^2 T}} \tag{18}$$

The likelihood of the DN route and the vanilla route arriving at the decision bound first to yield a choice i is, respectively:

$$L^{DN}(i,T) = \mathbf{g}(\mu_i^{DN}, T, \theta, \sigma)\mathbf{G}\left(\mu_j^{DN}, T, \theta, \sigma\right)\mathbf{G}(\mu_i, T, \theta, \sigma)\mathbf{G}\left(\mu_j, T, \theta, \sigma\right) \tag{19}$$

$$L(i,T) = \mathbf{g}(\mu_i, T, \theta, \sigma)\mathbf{G}\left(\mu_j, T, \theta, \sigma\right)\mathbf{G}(\mu_i^{DN}, T, \theta, \sigma)\mathbf{G}\left(\mu_j^{DN}, T, \theta, \sigma\right) \tag{20}$$

The final likelihood function of choice i at time T is thus: $L^{DN}(i,T) + L(i,T)$. These closed-form analytical solutions of choice probability and mean RT, as given by equations (9) and (10), match well with the model simulations performed in Chau et al. (2020)[18]. Empirically the trial-by-trial model predictions of a model with free σ (one for each route) vs. a model with σ = 1 are highly similar (across-participants mean Pearson's correlation coefficient $r(148)$ = .96 for ternary choice probabilities, $r(148)$ = .98 for binary choice probabilities, $r(148)$ = .97 for ternary RTs, $r(148)$ = .88 for binary RTs). Here we chose the simpler model with unit σ (lower BIC

compared to the free σ model: -169 for ternary trials, -790 for binary trials, summed across participants).

Static models

The static models were fit to trial-by-trial human choice probability (while ignoring RT). These models assume a softmax function that takes as inputs the options' utilities (AU, EV, or divisively normalised EV, see **Fig. 3**) and controls the level of decision noise via an inverse temperature parameter β:

$$p(\text{choice} = i) = \frac{e^{\beta U_i}}{\sum_i e^{\beta U_i}} \tag{21}$$

The model-predicted and the empirical choice probabilities were then used to calculate the binomial log-likelihood: $LL \propto p_e \log(p_m) + (1 - p_e)\log(1 - p_m)$, in which p_m and p_e denotes the model-predicted and empirical *p*(H over L), respectively.

Model fitting, comparison, and recovery

Models were fit to data via maximising the log-likelihood summed over trials. Because the *LL* functions all have analytical solutions, Matlab's fmincon was used for parameter optimisation. To avoid local optima, we refit each model to each participant's behavioural data at least 10 times using a grid of randomly generated starting values for the free parameters. Conversative criteria were used during the optimisation search: MaxFunEvals = MaxIter = 5,000; TolFun = TolX = 10^{-10}. We then calculated each model's posterior frequency and protected exceedance probability (i.e., the probability corrected for chance level that a model is more likely than any others in describing the data) using the variational Bayesian analysis (VBA) toolbox[78,79]. To be consistent with the methods in Chau et al.[18], we fit models to those trials with either H or L responses whilst ignoring a very small portion of trials in which D was accidentally chosen (median proportion of D choices: 4%, IQR: [2%, 8%], across participants).

Model comparison relied on cross-validation. That is, for each participant, we split the 150 trials into 5 folds and then fit each model to a 'training' set comprising 4 random folds of trials, obtained the best-fitting parameters, and finally calculated the *LL* summed across trials in the left-out 'test' fold. This process was repeated over test folds and the final cross-validated *LL* was computed as the mean *LL* across cross-validation folds. Model recovery was carried out by first simulating each model's behaviour (choice probability and RT) with its own best-fitting parameters obtained from fitting the model to each participant's raw data; then cross-fitting all models of interest to the simulated behaviour and calculating the *LL*s summed over trials; and finally comparing models' goodness-of-fit using Bayesian model comparison.

Subjective distortion of reward attributes

When modelling binary-choice behaviour, in addition to assuming linear subjective representations of reward attributes for simplicity, we also considered non-linear subjective functions[34]. The probability distortion undergoes a log-odds function (equation 22) that converts a probability value P to P_d, in a flexible manner (S-shape, inverted S-shape, concave, or convex, depending on parameterisation), whereas the magnitude distortion follows a power-law function (equation 23) that converts X to X_d:

$$\log\left(\frac{P_d}{1 - P_d}\right) = \eta \log\left(\frac{P}{1 - P}\right) + (1 - \eta)\log\left(\frac{P_0}{1 - P_0}\right) \tag{22}$$

$$X_d = X^\gamma \tag{23}$$

Where $\eta > 0$, P_0 ($0 < P_0 < 1$), and $\gamma > 0$ are the 3 parameters controlling the extent of non-linearity of the distortion functions.

In addition to the log-odds function, we also considered another non-linear function (only 1 free parameter τ) for the probability distortion, described by equation 24:

$$P_d = \frac{P^\tau}{[P^\tau + (1-P)^\tau]^{\frac{1}{\tau}}} \tag{24}$$

This particular function underlies the subjective probability distortion in a model of the *prospect theory*[35]. More specifically, the prospect-theory model assumes a multiplicative utility that multiplies X_d (equation 23) with P_d (equation 24). Another popular multiplicative model is the standard *expected utility*[20] in which the probability function is linear (i.e., τ is fixed to 1) while γ governing X_d is a free parameter. We included these 2 additional multiplicative models when comparing additive vs. multiplicative strategies (**Figure 3—figure supplement 1**).

Condition-nonspecific response bias correction

Choice behaviour can differ between a binary condition and a ternary condition, simply due to the increased number of options (higher cognitive demands). To estimate these condition-nonspecific biases, we used a permutation approach to shuffling the mappings between the ternary conditions and their matched binary conditions (5,000 permutations). This yielded a mean bias term for the 'T minus matched B' data, and was eventually removed from the re-referenced ternary performance ('T – B'), for relative accuracy and for RT, respectively.

Generalised linear model (GLM) analysis of relative choice accuracy

We used GLM approach to analysing choice behaviour. For ternary conditions, we focused on the *relative* accuracy, i.e., the proportion of H choices among trials with H or L choices (while ignoring a small number of D-choice trials), which has been commonly used as a measure of the violation of the Independence from Irrelevant Alternatives (IIA) principle. Matched binary trials were identified for each ternary trial (see **Figure 2—figure supplement 1**). Among the 150 ternary trials there are 149 unique ternary conditions with a unique combination of probability (P) and magnitude (X) attributes across the 3 alternatives (H, L, and D), whereas among the 150 binary trials there are only 95 unique conditions with a specific combination of P and X across the 2 available options (H and L). The ternary and binary trials do not have 'one-to-one' mapping. Some ternary trials have more than 1 matched binary trial. The different counts of matched binary trials were explicitly considered as 'observation weights' in the GLM. We used logit binomial GLM to analyse trial-by-trial relative choice accuracy (MATLAB's glmfit with 'binomial' distribution and 'logit' link function). The GLMs in **Fig. 2** (**d** and **e**; using relative distractor variable 'DV – HV') and **Figure 2—figure supplement 2** (using absolute distractor variable 'DV') were specified as follows:

Logit(p(H over L)) ~ $\beta_0 + \beta_1$ z(HV – LV) + β_2 z(DV – HV) + β_3 z(HV – LV) x z(DV – HV)
Logit(p(H over L)) ~ $\beta_0 + \beta_1$ z(HV – LV) + β_2 z(DV) + β_3 z(HV – LV) x z(DV)

z(): z-score normalisation of the regressor vector. The interaction terms were constructed *after* individual component terms were z-scored[19]. The analysis shown in **Fig. 2f** combined ternary (T) and binary (B) relative accuracies in a single GLM and assessed how distractor-related effects interacted with 'Condition' (i.e., 'C', a dummy variable: 0 for B, 1 for T). The code for reproducing Fig. 2 and Figure 2—figure supplement 2 and 3 GLMs can be found here: https://github.com/YinanCao/multiattribute-distractor/

Construction of binned maps

The choice-accuracy and RT maps (**Fig. 2b**, **c; Fig. 3a**) were constructed using the exact binning approach of Chau et al. (2020)[18]. Data were averaged using a sliding square window inside the space defined by the 2 expected-value differences, (DV – HV) and (HV – LV). The edge length of the window is 30% quantile difference along each dimension, with step-size being equal to 1% quantile difference. This binning approach smooths the data and results in a map of size 71-by-71.

Acknowledgements

We thank the authors of the original studies for sharing their data publicly (Chau et al., 2014[4]: https://doi.org/10.5061/dryad.040h9t7; Gluth et al., 2018[19]: https://osf.io/8r4fh/). This work was supported by the EU Horizon 2020 Research and Innovation Program (ERC starting grant no. 802905) to K.T. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Contributions

Y.C. performed the modelling and analyses. K.T. supervised all stages of the work. Y.C. and K.T. wrote the paper.

Code availability

https://github.com/YinanCao/multiattribute-distractor/

Competing interests

The authors declare no competing interests.

References

1. Stone, L. S. & Thompson, P. Human speed perception is contrast dependent. *Vision Res.* **32**, 1535–1549 (1992).
2. Summerfield, C. & Tsetsos, K. Do humans make good decisions? *Trends Cogn. Sci.* **19**, 27–34 (2015).
3. Huber, J., Payne, J. W. & Puto, C. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *J. Consum. Res.* **9**, 90–98 (1982).
4. Chau, B. K. H., Kolling, N., Hunt, L. T., Walton, M. E. & Rushworth, M. F. S. A neural mechanism underlying failure of optimal choice with multiple alternatives. *Nat. Neurosci.* **17**, 463–470 (2014).
5. Louie, K., Khaw, M. W. & Glimcher, P. W. Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 6139–6144 (2013).
6. Dumbalska, T., Li, V., Tsetsos, K. & Summerfield, C. A map of decoy influence in human multialternative choice. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25169–25178 (2020).
7. Luce, R. D. The choice axiom after twenty years. *J. Math. Psychol.* **15**, 215–233 (1977).
8. Rieskamp, J., Busemeyer, J. R. & Mellers, B. A. Extending the bounds of rationality: Evidence and theories of preferential choice. *J. Econ. Lit.* **44**, 631–661 (2006).
9. Vlaev, I., Chater, N., Stewart, N. & Brown, G. D. A. Does the brain calculate value? *Trends Cogn. Sci.* **15**, 546–554 (2011).
10. Busemeyer, J. R., Gluth, S., Rieskamp, J. & Turner, B. M. Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends Cogn. Sci.* **23**, 251–263 (2019).
11. Hunt, L. T., Dolan, R. J. & Behrens, T. E. J. Hierarchical competitions subserving multi-attribute choice. *Nat. Neurosci.* **17**, 1613–1622 (2014).
12. Tsetsos, K., Usher, M. & Chater, N. Preference reversal in multiattribute choice. *Psychol. Rev.* **117**, 1275–1293 (2010).
13. Turner, B. M., Schley, D. R., Muller, C. & Tsetsos, K. Competing theories of multialternative, multiattribute preferential choice. *Psychol. Rev.* **125**, 329–362 (2018).
14. Usher, M. & McClelland, J. L. Loss aversion and inhibition in dynamical models of multialternative choice. *Psychol. Rev.* **111**, 757–769 (2004).
15. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. 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).

16. Gluth, S., Kern, N., Kortmann, M. & Vitali, C. L. Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nat Hum Behav* **4**, 634–645 (2020).
17. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2011).
18. Chau, B. K., Law, C.-K., Lopez-Persem, A., Klein-Flügge, M. C. & Rushworth, M. F. Consistent patterns of distractor effects during decision making. *Elife* **9**, (2020).
19. Gluth, S., Spektor, M. S. & Rieskamp, J. Value-based attentional capture affects multi-alternative decision making. *Elife* **7**, (2018).
20. von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior*. (Princeton University Press, 2007).
21. Bernoulli, D. Exposition of a New Theory on the Measurement of Risk. *Econometrica* **22**, 23–36 (1954).
22. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).
23. Pratt. Risk aversion in the small and in the large. *Uncertainty in economics* (1978).
24. Weber, E. U., Shafir, S. & Blais, A.-R. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychol. Rev.* **111**, 430–445 (2004).
25. Rouault, M., Drugowitsch, J. & Koechlin, E. Prefrontal mechanisms combining rewards and beliefs in human decision-making. *Nat. Commun.* **10**, 301 (2019).
26. Stewart, N. Information integration in risky choice: identification and stability. *Front. Psychol.* **2**, 301 (2011).
27. Farashahi, S., Donahue, C. H., Hayden, B. Y., Lee, D. & Soltani, A. Flexible combination of reward information across primates. *Nat Hum Behav* **3**, 1215–1224 (2019).
28. Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
29. Lu, Z.-L. & Dosher, B. A. Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychol. Rev.* **115**, 44–82 (2008).
30. Bongioanni, A. *et al.* Activation and disruption of a neural mechanism for novel choice in monkeys. *Nature* **591**, 270–274 (2021).
31. Tsetsos, K. *et al.* Economic irrationality is optimal during noisy decision making. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3102–3107 (2016).
32. Hayden, B. Y. & Niv, Y. The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *Behav. Neurosci.* **135**, 192–201 (2021).
33. Pirrone, A., Reina, A., Stafford, T., Marshall, J. A. R. & Gobet, F. Magnitude-sensitivity: rethinking decision-making. *Trends Cogn. Sci.* **26**, 66–80 (2022).
34. Zhang, H. & Maloney, L. T. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front. Neurosci.* **6**, 1 (2012).
35. Kahneman, D. & Tversky, A. Prospect Theory: An analysis of decisions under risk. *Econometrica* **47**, 278 (1979).
36. Spektor, M. S., Bhatia, S. & Gluth, S. The elusiveness of context effects in decision making. *Trends Cogn. Sci.* **25**, 843–854 (2021).
37. Li, V., Michael, E., Balaguer, J., Castañón, S. H. & Summerfield, C. Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making. *Proceedings of the National Academy of Sciences* **115**, E8825–E8834 (2018).
38. Wyart, V. & Koechlin, E. Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences* **11**, 109–115 (2016).
39. Webb, R., Glimcher, P. W. & Louie, K. Divisive normalization does influence decisions with multiple alternatives. *Nat Hum Behav* (2020) doi:10.1038/s41562-020-00941-5.
40. Gluth, S., Kern, N. & Vitali, C. L. Reply to: Divisive normalization does influence decisions with multiple alternatives. *Nat Hum Behav* (2020) doi:10.1038/s41562-020-00942-4.

41. Pettibone, J. C. & Wedell, D. H. Testing alternative explanations of phantom decoy effects. *J. Behav. Decis. Mak.* **20**, 323–341 (2007).
42. Farmer, G. D., Warren, P. A., El-Deredy, W. & Howes, A. The effect of expected value on attraction effect preference reversals. *J. Behav. Decis. Mak.* **30**, 785–793 (2017).
43. Juechems, K. & Summerfield, C. Where Does Value Come From? *Trends Cogn. Sci.* **23**, 836–850 (2019).
44. Fellows, L. K. Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. *Brain* **129**, 944–952 (2006).
45. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
46. Bellman, R. E. *Dynamic programming*. (Princeton University Press, 1957).
47. Niv, Y. Learning task-state representations. *Nat. Neurosci.* **22**, 1544–1553 (2019).
48. Luyckx, F., Spitzer, B., Blangero, A., Tsetsos, K. & Summerfield, C. Selective integration during sequential sampling in posterior neural signals. *Cereb. Cortex* **30**, 4454–4464 (2020).
49. Usher, M., Tsetsos, K., Glickman, M. & Chater, N. Selective integration: An attentional theory of choice biases and adaptive choice. *Curr. Dir. Psychol. Sci.* **28**, 552–559 (2019).
50. Noguchi, T. & Stewart, N. Multialternative decision by sampling: A model of decision making constrained by process data. *Psychol. Rev.* **125**, 512–544 (2018).
51. Trueblood, J. S., Brown, S. D. & Heathcote, A. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychol. Rev.* **121**, 179–205 (2014).
52. Bhatia, S. Associations and the accumulation of preference. *Psychol. Rev.* **120**, 522–543 (2013).
53. Roe, R. M., Busemeyer, J. R. & Townsend, J. T. Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychol. Rev.* **108**, 370–392 (2001).
54. Koechlin, E. Human Decision-Making beyond the Rational Decision Theory. *Trends in cognitive sciences* vol. 24 4–6 (2020).
55. Tversky, A. & Wakker, P. Risk Attitudes and Decision Weights. *Econometrica* **63**, 1255–1280 (1995).
56. Birnbaum, M. H. Evidence against prospect theories in gambles with positive, negative, and mixed consequences. *J. Econ. Psychol.* **27**, 737–761 (2006).
57. Birnbaum, M. H. New paradoxes of risky decision making. *Psychol. Rev.* **115**, 463–501 (2008).
58. Tom, S. M., Fox, C. R., Trepel, C. & Poldrack, R. A. The neural basis of loss aversion in decision-making under risk. *Science* **315**, 515–518 (2007).
59. Chater, N., Johansson, P. & Hall, L. The non-existence of risk attitude. *Front. Psychol.* **2**, 303 (2011).
60. De Martino, B., Kumaran, D., Seymour, B. & Dolan, R. J. Frames, biases, and rational decision-making in the human brain. *Science* **313**, 684–687 (2006).
61. Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
62. Kennerley, S. W., Dahmubed, A. F., Lara, A. H. & Wallis, J. D. Neurons in the frontal lobe encode the value of multiple decision variables. *J. Cogn. Neurosci.* **21**, 1162–1178 (2009).
63. Koechlin, E. An evolutionary computational theory of prefrontal executive function in decision-making. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, (2014).
64. Yamada, H., Imaizumi, Y. & Matsumoto, M. Neural Population Dynamics Underlying Expected Value Computation. *J. Neurosci.* **41**, 1684–1698 (2021).
65. Suzuki, S., Cross, L. & O'Doherty, J. P. Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nat. Neurosci.* **20**, 1780–1786 (2017).
66. Cao, Y., Summerfield, C., Park, H., Giordano, B. L. & Kayser, C. Causal Inference in the Multisensory Brain. *Neuron* **102**, 1076-1087.e8 (2019).
67. Williams, T. B. *et al.* Testing models at the neural level reveals how the brain computes subjective value. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
68. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).

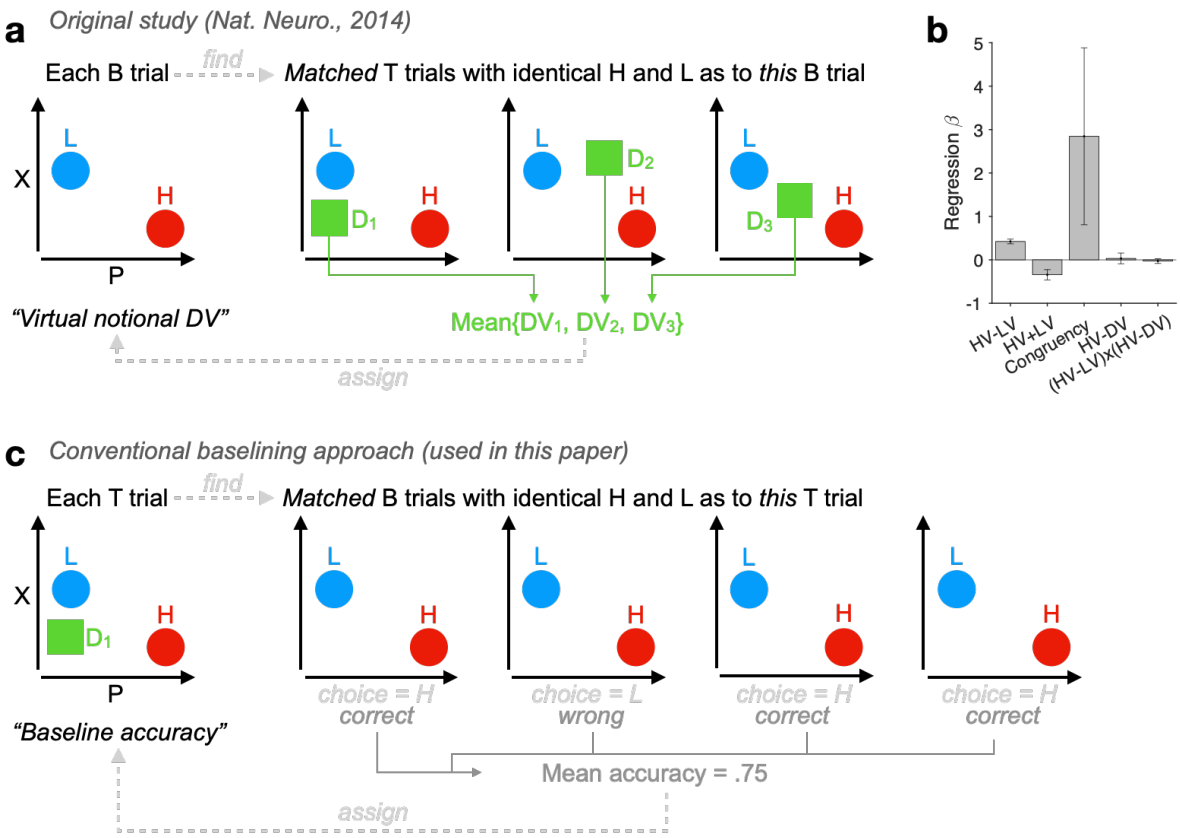
1045 69. Roesch, M. R., Taylor, A. R. & Schoenbaum, G. Encoding of time-discounted rewards in
1046 orbitofrontal cortex is independent of value representation. *Neuron* **51**, 509–520 (2006).
1047 70. Soltani, A., De Martino, B. & Camerer, C. A range-normalization model of context-dependent
1048 choice: a new model and evidence. *PLoS Comput. Biol.* **8**, e1002607 (2012).
1049 71. Kable, J. W. & Glimcher, P. W. The neurobiology of decision: consensus and controversy. *Neuron*
1050 **63**, 733–745 (2009).
1051 72. Levy, D. J. & Glimcher, P. W. The root of all value: a neural common currency for choice. *Curr. Opin.*
1052 *Neurobiol.* **22**, 1027–1038 (2012).
1053 73. Padoa-Schioppa, C. & Assad, J. A. The representation of economic value in the orbitofrontal cortex
1054 is invariant for changes of menu. *Nat. Neurosci.* **11**, 95–102 (2008).
1055 74. Mazurek, M. E., Roitman, J. D., Ditterich, J. & Shadlen, M. N. A role for neural integrators in
1056 perceptual decision making. *Cereb. Cortex* **13**, 1257–1269 (2003).
1057 75. Usher, M., Olami, Z. & McClelland, J. L. Hick's law in a stochastic race model with speed–accuracy
1058 tradeoff. *J. Math. Psychol.* **46**, 704–715 (2002).
1059 76. Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E. & Pouget, A. Optimal multisensory
1060 decision-making in a reaction-time task. *Elife* **3**, (2014).
1061 77. Palmer, J., Huk, A. C. & Shadlen, M. N. The effect of stimulus strength on the speed and accuracy
1062 of a perceptual decision. *J. Vis.* **5**, 376–404 (2005).
1063 78. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—
1064 revisited. *Neuroimage* **84**, 971–985 (2014).
1065 79. Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for
1066 neurobiological and behavioural data. *PLoS Comput. Biol.* **10**, e1003441 (2014).

Supplemental Information

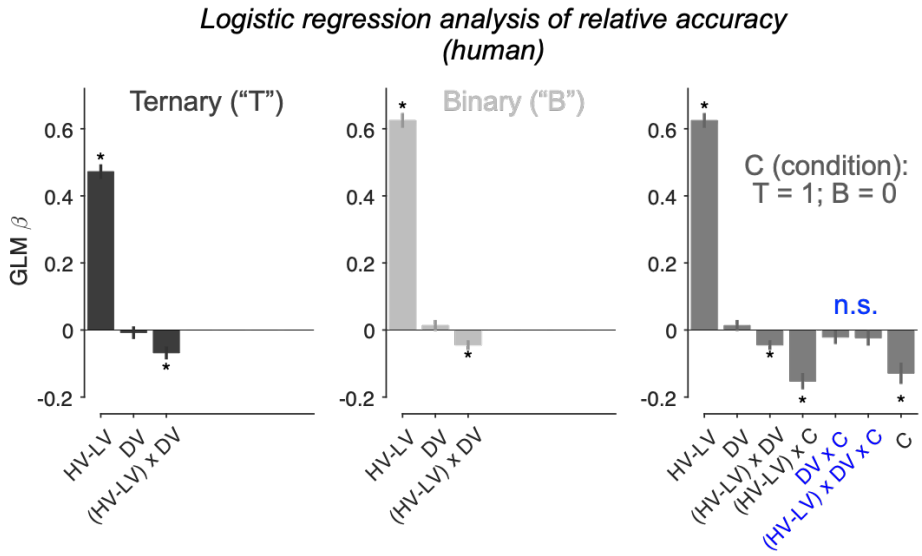
Supplemental Information includes 6 figures and 4 tables.

Supplemental Figures

Figure 2—figure supplement 1 (related to Fig. 2). Distinctions between our baselining approach and the method in the original study[4]. (**a**) Method in the original study[4]: A schematic showing that the mean DV across relevant ternary trials ('T') is taken as a notional DV for each individual binary trial ('B'). For each participant, among the 150 T trials there are 149 unique T conditions with a specific combination of P and X attributes across the 3 alternatives, while among the 150 B trials there are only 95 unique B conditions with a specific combination of P and X across the 2 options. The T and B trials do not have one-to-one mapping even if the unique stimulus positions (among 4 quadrants of the screen) are taken into consideration. (**b**) A successful reproduction of an invalid control analysis in the *Supplemental Fig. 3b* of the original study[4]: *"The effect of HV-D on human behavior was specific to distractor trials"* using the notional DV depicted in panel a. (**c**) Our baselining approach uses a new dependent variable, the trial-by-trial baseline accuracy, via identifying all matched B trials (number of trials 1<= nB <=6) of every T trial. Of note, when analysing choice accuracies in B trials using logistic regression (e.g., Fig. 2), Matlab's glmfit takes a 2-column response matrix whose first column indicates the number of correct responses (H choices) for each observation and second column indicates nB for each observation.



1090 **Figure 2—figure supplement 2** (related to Fig. 2). Re-assessing distractor effects beyond target-
1091 related covariations. The only difference between this figure and Fig. 2 in the main text is that
1092 the regressor DV – HV in Fig. 2 is replaced by DV here. Error bars = ± s.e.m. (*N* = 144
1093 participants).
1094



1095

Figure 2—figure supplement 3 (related to Fig. 2). Re-assessing distractor effects beyond target-related covariations. The GLMs here included an additional regressor HV + LV. Error bars = ± s.e.m. (N = 144 participants). Asterisks: significant effects (p < .05; 2-sided 1-sample t-tests of GLM beta coefficients against 0) following Holm's sequential Bonferroni correction for multiple comparisons.

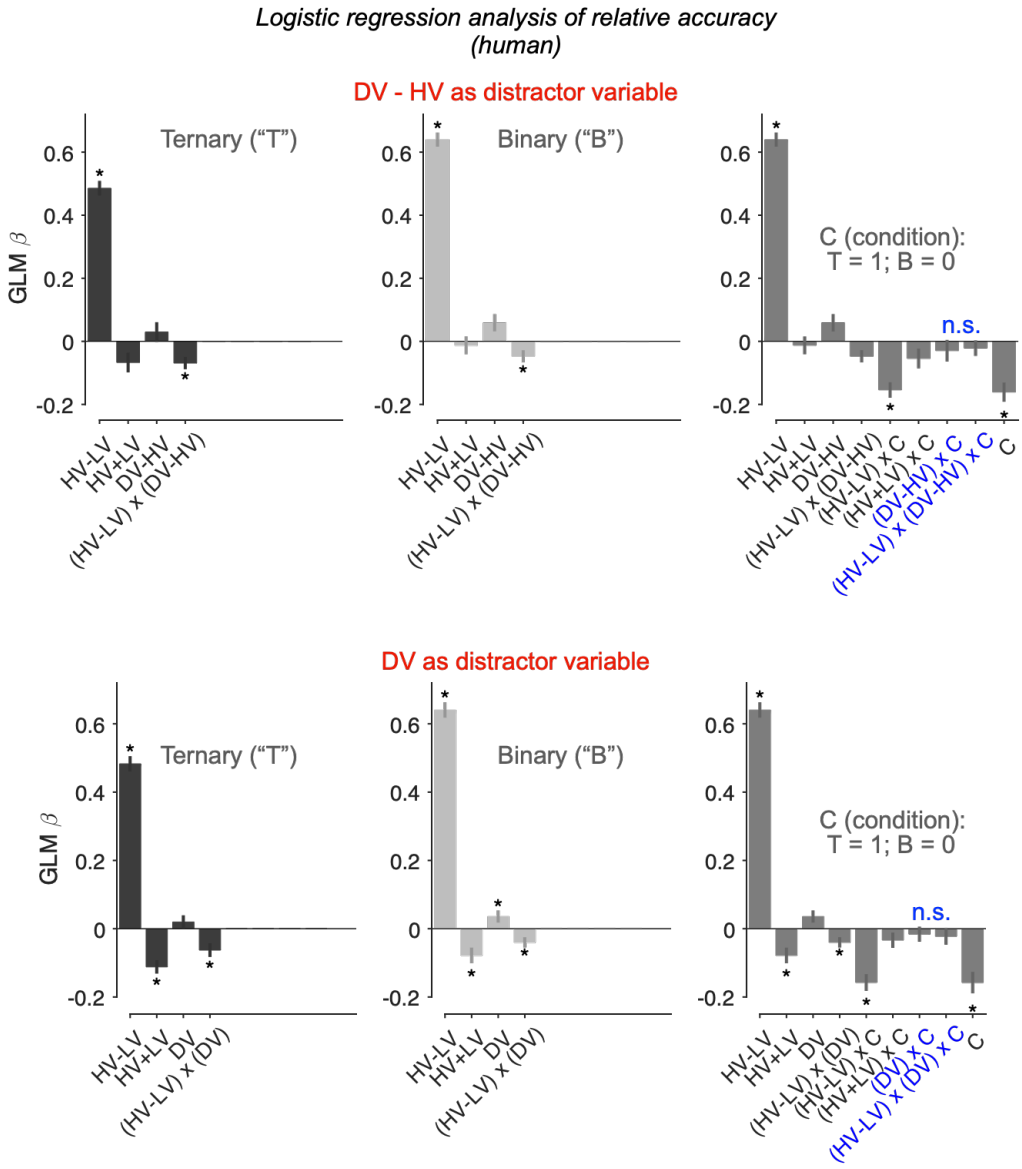


Figure 3—figure supplement 1 (related to Fig. 3). (**a**) Additional multiplicative models (expected utility and prospect theory) fit to binary-choice trials. (**b**) Visualisation of best-fitting utility and probability functions of the prospect-theory model. Gray curves: individual participants' best-fitting functions; red curves: median across participants (*N* = 144). The typical best-fitting weighting function tends to overweigh low probabilities and the typical utility function (distorted magnitude) tends to be concave (i.e., compressive non-linearity).

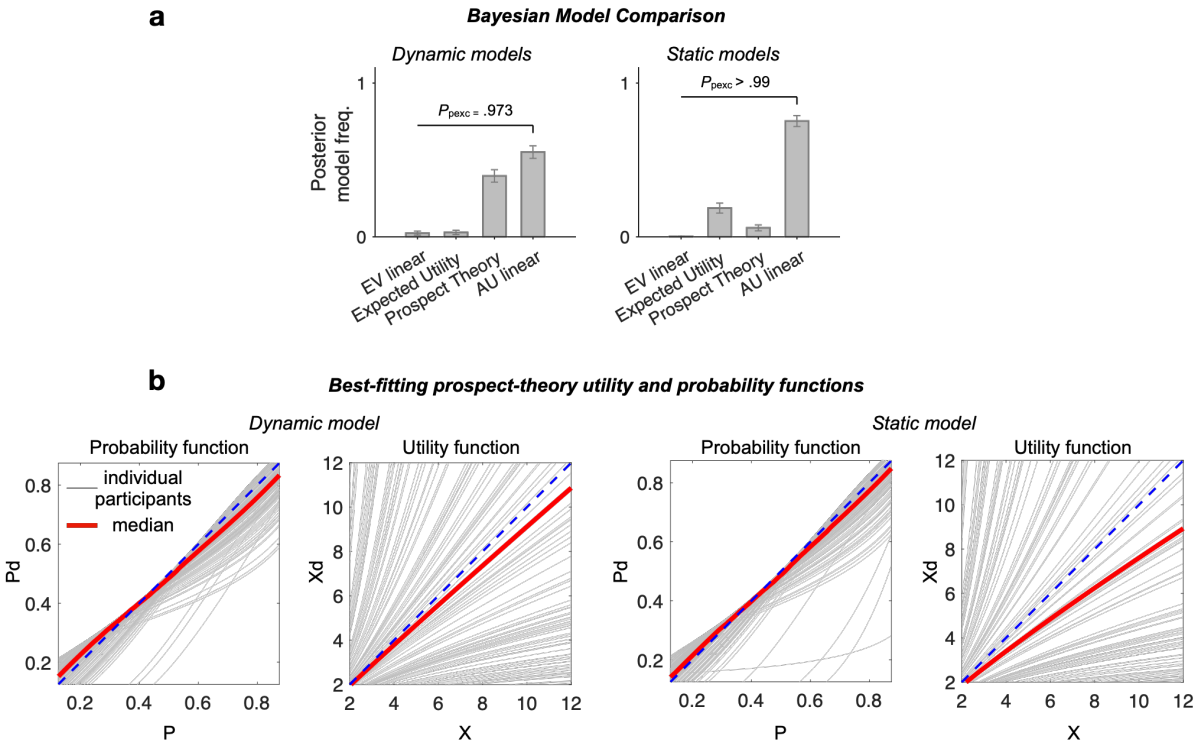


Figure 4—figure supplement 1 (related to Fig. 4). (**a**) Bayesian random-effects model comparison based on cross-validated log-likelihood (5 folds). SI: Selective integration model; Adap. Gain: adaptive gain model; CI: context independent; EV: expected value; AU: additive utility. (**b**, **c**) Dual-route and divisive normalisation (DN) model predictions for the relative accuracy change (Ternary minus Binary, T – B) as a function of (**b**) both D's AU dominance (colour of the bars) and D's position similarity to H and L in the 2D attribute space, or (**c**) D's AU dominance only (main effect). AU: additive utility. EV: expected value.

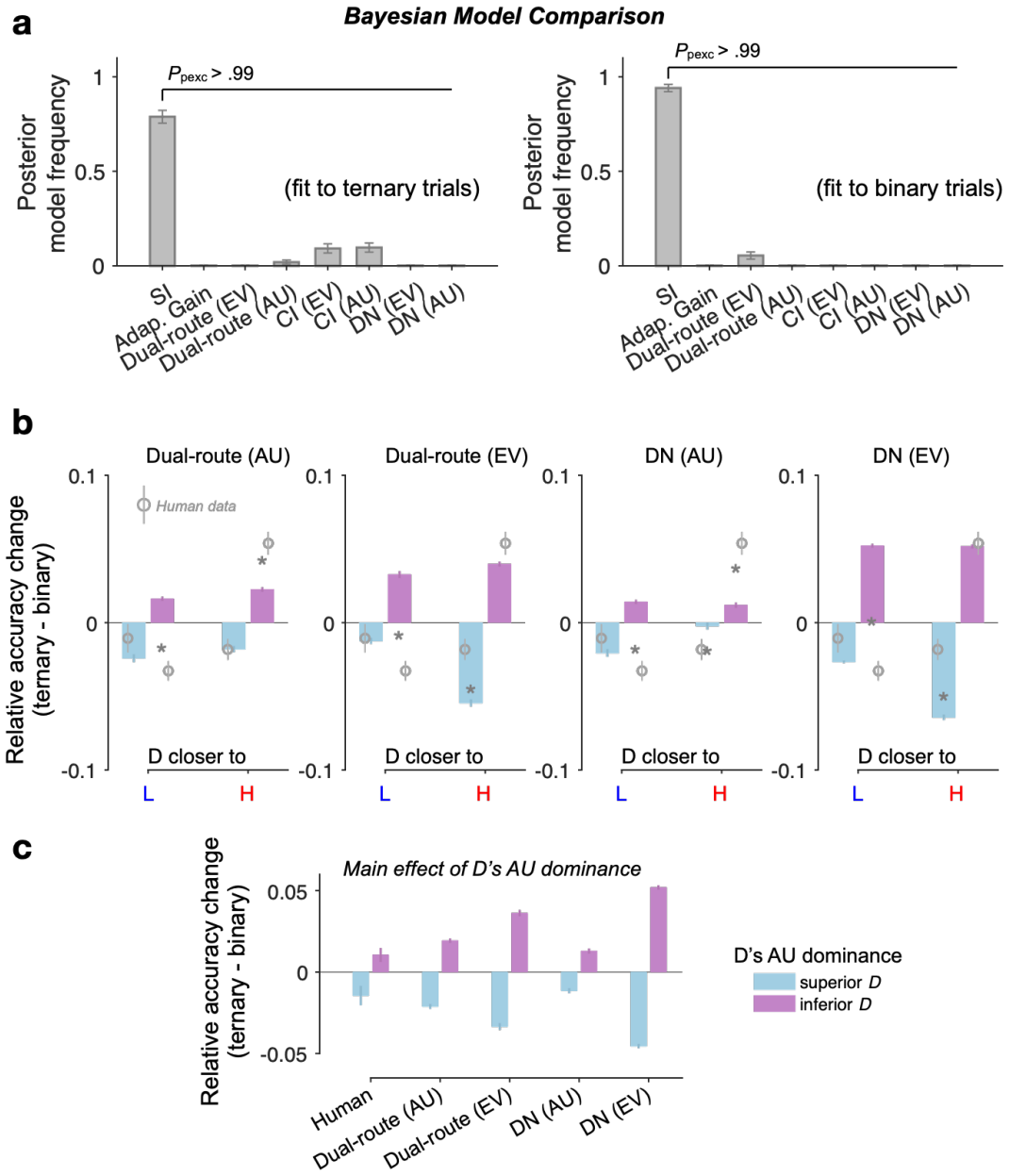
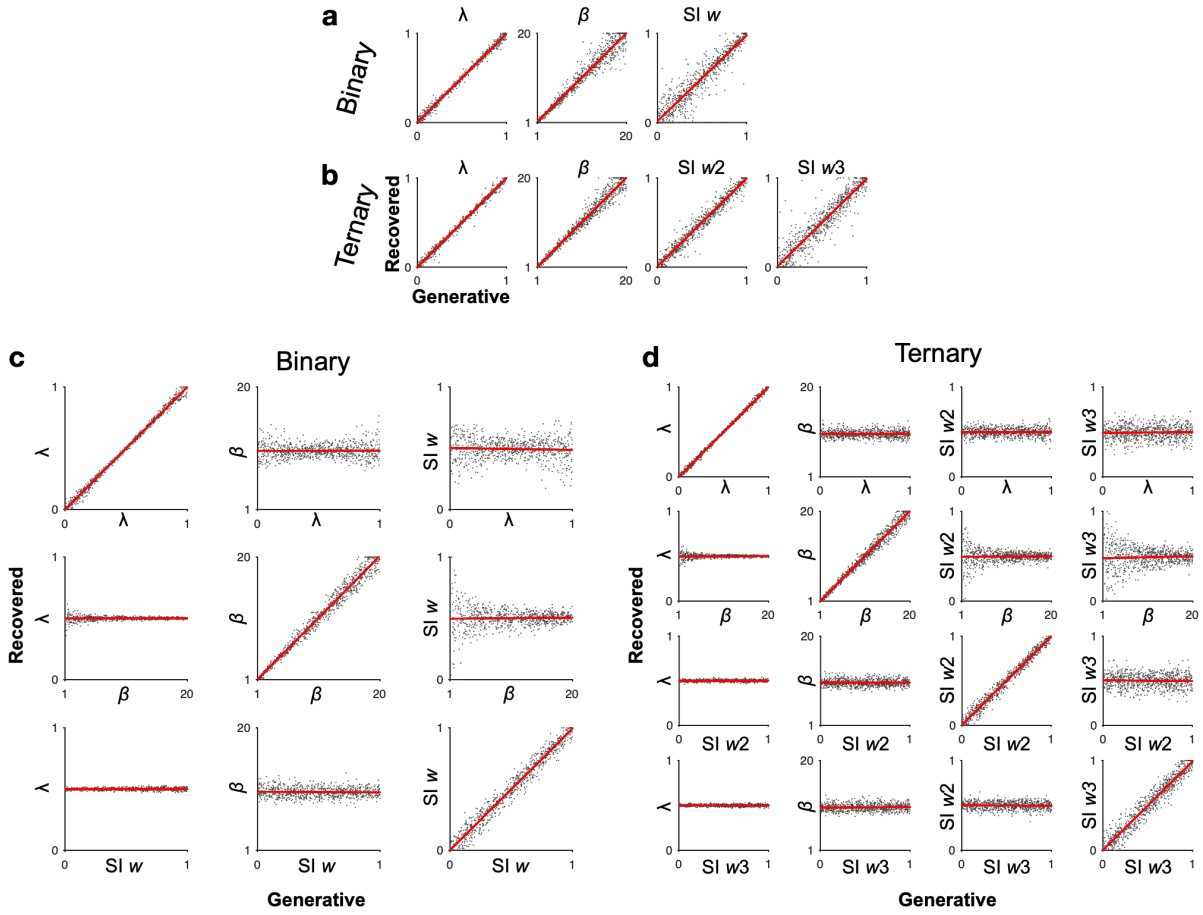


Figure 5—figure supplement 1 (related to Fig. 5). Parameter recovery. (**a**, **b**) Recovery of SI model parameters by simulating choice probabilities (1,000 trials) based on a grid of uniformly distributed parameters' values (n = 800 each) and then estimating the best-fitting parameters. (**c**, **d**) Each row shows a case of simulating choice probabilities by varying 1 model parameter while fixing all other parameters (λ = 0.5, β = 10, SI w = 0.5) and then estimating the best-fitting parameter values via model fitting. On-diagonal panels show the recovery of each parameter as a function of itself (all |Pearson's r| > .99), whilst off-diagonal panels suggest that variation in irrelevant parameters do not bias the recovery of any given parameter of interest (all |Pearson's r| < .056, p > .1).



Supplemental Tables

Supplemental Table 1 (related to Fig. 3). Optimal parameter estimates of dynamic models: Mean (SE) and cross-validated log-likelihood (CV LL).

AU FFI linear (n = 6 free parameters)

λ	k	θ	I_0	t_{nd}	c	CV LL
.475 (.017)	3.89 (.11)	3.21 (.097)	2.81 (.06)	.141 (.012)	.76 (.011)	-958.5

EV FFI linear (n = 5)

k	θ	I_0	t_{nd}	c	CV LL
3.38 (.097)	3.076 (.09)	3.14 (.061)	.152 (.012)	.85 (.01)	-1191.3

EV (+ DN) FFI linear (n = 5)

k	θ	I_0	t_{nd}	c	CV LL
1.96 (.061)	2.92 (.097)	2.59 (.07)	.167 (.012)	.38 (.009)	-1286.1

Dual-route EV linear (n = 6)

k	k^{DN}	θ	I_0	t_{nd}	f_{MI}	CV LL
5.62 (.21)	2.53 (.11)	2.77 (.08)	1.99 (.064)	.169 (.011)	.75 (.021)	-1150

AU FFI non-linear (n = 9)

λ	k	θ	I_0	t_{nd}	c	η	P_0	γ	CV LL
.337	11.7	3.26	1.29	.144	.77	5.6	.69	4.8	-822
(.023)	(.61)	(.099)	(.17)	(.012)	(.011)	(1.04)	(.032)	(.88)	

EV FFI non-linear (n = 8)

k	θ	I_0	t_{nd}	c	η	P_0	γ	CV LL
6.25	3.22	2.48	.15	.81	2.39	.56	1.04	-897
(.41)	(.09)	(.16)	(.012)	(.012)	(.57)	(.032)	(.14)	

EV (+ DN) FFI non-linear (n = 8)

k	θ	I_0	t_{nd}	c	η	P_0	γ	CV LL
3.52	3.07	2.67	.16	.45	17.98	.61	13.8	-1052
(.44)	(.096)	(.083)	(.012)	(.015)	(1.7)	(.03)	(1.4)	

Dual-route EV non-linear (n = 9)

k	k^{DN}	θ	I_0	t_{nd}	f_{MI}	η	P_0	γ	CV LL
6.58	4.29	2.95	1.69	.16	.67	9.23	.56	5.74	-927
(.38)	(.40)	(.09)	(.12)	(.011)	(.027)	(1.28)	(.03)	(.96)	

Supplemental Table 2 (related to Figs. 2, 3). Optimal parameter estimates of static models: Mean (SE), and cross-validated log-likelihood (CV LL).

AU softmax linear (n = 2 free parameters)

λ	β	CV LL
.457 (.017)	9.15 (.302)	-2330.1

EV softmax linear (n = 1)

β	CV LL
6.79 (.198)	-2645.5

EV (+ DN) softmax linear (n = 1)

β	CV LL
3.35 (.11)	-2656.4

AU softmax non-linear (n = 5)

λ	β	η	P_0	γ	CV LL
.327 (.019)	16.4 (.47)	4.72 (.938)	.669 (.031)	4.43 (.792)	-2193.3

EV softmax non-linear (n = 4)

β	η	P_0	γ	CV LL
33.9 (8.81)	3.69 (.74)	.524 (.036)	.82 (.13)	-2388.8

EV (+ DN) softmax non-linear (n = 4)

β	η	P_0	γ	CV LL
31.3 (6.53)	13.5 (1.62)	.607 (.031)	9.44 (1.24)	-2481.2

Supplemental Table 3 (related to Fig. 2). GLMs using subjective additive-utility (AU) regressors. * Significant effects ($p < .05$; 2-sided 1-sample t-tests of GLM coefficients against 0) following Holm's sequential Bonferroni correction for multiple comparisons. C: 'Condition' (binary vs. ternary).

AU_D as (absolute) distractor variable (with 'AU sum' covariate):

	T value	p-value (uncorrected)
$AU_H - AU_L$	12.3	< .0001*
$AU_H + AU_L$	11.9	< .0001*
AU_D	2.37	.0193
$(AU_H - AU_L)$ x AU_D	-.099	.92
$(AU_H - AU_L)$ x C	-6.09	< .0001*
$(AU_H + AU_L)$ x C	-5.24	< .0001*
AU_D x C	-2.27	.024
$(AU_H - AU_L)$ x AU_D x C	-1.64	.102
C	-5.04	< .0001*

AU_D as (absolute) distractor variable (without 'AU sum' covariate):

	T value	p-value (uncorrected)
$AU_H - AU_L$	13.9	< .0001*
AU_D	3.93	< .001*
$(AU_H - AU_L)$ x AU_D	-.44	.66
$(AU_H - AU_L)$ x C	-6.56	< .0001*
AU_D x C	-2.87	.0047*
$(AU_H - AU_L)$ x AU_D x C	-1.74	.083
C	-8.06	< .0001*

$AU_D - AU_H$ as (relative) distractor variable (with 'AU sum' covariate):

	T value	p-value (uncorrected)
$AU_H - AU_L$	13.1	< .0001*
$AU_H + AU_L$	11.7	< .0001*
$AU_D - AU_H$	-.35	.73
$(AU_H - AU_L)$ x $(AU_D - AU_H)$	-4.25	< .0001*
$(AU_H - AU_L)$ x C	-6.24	< .0001*
$(AU_H + AU_L)$ x C	-5.63	< .0001*
$(AU_D - AU_H)$ x C	-1.86	.065
$(AU_H - AU_L)$ x $(AU_D - AU_H)$ x C	-.54	.59
C	-5.39	< .0001*

$AU_D - AU_H$ as (relative) distractor variable (without 'AU sum' covariate):

	T value	p-value (uncorrected)
$AU_H - AU_L$	13.7	< .0001*
$AU_D - AU_H$	-10.2	< .0001*
$(AU_H - AU_L)$ x $(AU_D - AU_H)$	-5.65	< .0001*
$(AU_H - AU_L)$ x C	-6.41	< .0001*
$(AU_D - AU_H)$ x C	1.95	.054
$(AU_H - AU_L)$ x $(AU_D - AU_H)$ x C	.22	.83
C	-7.18	< .0001*

Supplemental Table 4 (related to Fig. 4). Optimal parameter estimates of context-dependent dynamic models: Mean (SE), and cross-validated log-likelihood (CV LL).

SI AU FFI (n = 8 free parameters) [Ternary]

λ	SI w_2	SI w_3	k	θ	I_0	t_{nd}	c	CV LL
.534	.25	.43	2.44	4.73	3.82	.024	.786	-1310.5
(.019)	(.029)	(.034)	(.078)	(.108)	(.074)	(.006)	(.02)	

SI AU FFI (n = 7) [Binary]

λ	SI w_2	k	θ	I_0	t_{nd}	c	CV LL
.487	.31	3.54	3.21	2.91	.145	.74	-915
(.016)	(.02)	(.11)	(.097)	(.061)	(.012)	(.015)	

Adaptive Gain AU FFI (n = 10) [Ternary]

λ	b_p	s_p	b_x	s_x	k	θ	I_0	t_{nd}	c	CV LL
.43	-.24	.23	.033	.24	44.2	4.7	.54	.023	.56	-1530
(.027)	(.12)	(.04)	(.12)	(.056)	(5.84)	(.11)	(.44)	(.006)	(.034)	

Adaptive Gain AU FFI (n = 10) [Binary]

λ	b_p	s_p	b_x	s_x	k	θ	I_0	t_{nd}	c	CV LL
.42	-.16	.15	-.092	.18	16.3	3.16	-2.68	.15	.19	-1012
(.023)	(.039)	(.02)	(.074)	(.051)	(2.4)	(.09)	(.43)	(.012)	(.02)	

Dual-route EV (n = 6) [Ternary]

k	k^{DN}	θ	I_0	t_{nd}	f_{MI}	CV LL
3.21 (.15)	1.68 (.09)	4.16 (.097)	2.85 (.065)	.035 (.008)	.72 (.028)	-1579

Dual-route EV (n = 6) [Binary]

k	k^{DN}	θ	I_0	t_{nd}	f_{MI}	CV LL
5.62 (.21)	2.53 (.11)	2.77 (.08)	1.99 (.064)	.169 (.011)	.75 (.021)	-1150