Preference Reversal in Multiattribute Choice

Konstantinos Tsetsos
University College London

Marius Usher
Tel Aviv University and Birkbeck College

Nick Chater
University College London

A central puzzle for theories of choice is that people’s preferences between options can be reversed by the presence of decoy options (that are not chosen) or by the presence of other irrelevant options added to the choice set. Three types of reversal effect reported in the decision-making literature, the attraction, compromise, and similarity effects, have been explained by a number of theoretical proposals. Yet a major theoretical challenge is capturing all 3 effects simultaneously. We review the range of mechanisms that have been proposed to account for decoy effects and analyze in detail 2 computational models, decision field theory (Roe, Busemeyer, & Townsend, 2001) and leaky competing accumulators (Usher & McClelland, 2004), that aim to combine several such mechanisms into an integrated account. By simulating the models, we examine differences in the ways the decoy effects are predicted. We argue that the LCA framework, which follows on Tversky’s relational evaluation with loss aversion (Tversky & Kahneman, 1991), provides a more robust account, suggesting that common mechanisms are involved in both high-level decision making and perceptual choice, for which LCA was originally developed.

**Keywords:** decision making, decoy effects, computational modes, dynamic models, loss aversion

Confronted with an unusually short dessert menu, Ms. X vacillates between two options, A and B. Finally, she plumps for A, at which point the waiter responds that, in fact, there is also the daily special, Option C. “Thank goodness you told me that,” says Ms. X, relieved, “In that case, I’d prefer B.” There is something paradoxical about Ms. X’s change of heart. How can the availability of a third option, C, possibly affect whether A or B is preferred? The relative pleasure of eating Dessert A or B surely should depend on the properties of A and B alone and not on the properties of any other dessert C, whether that C is an available option or not. To hammer home how paradoxical any influence of C might be, let us push the story a little further. The waiter returns with Dessert B and says, “Actually, the chef has just told me that C is sold out.” “In that case, I’d like to switch back to A, please,” decides Ms. X.

The puzzling behavior of Ms. X in this situation is a case of contextual preference reversal. It is fascinating that such reversals have been reported to characterize human decision making between alternatives that vary on several dimensions, as illustrated in Figure 1, where one has to choose one out of several cars that vary on two attributes (i.e., economy and quality). Three such reversal effects have been reported in the literature. The most puzzling of them are the attraction effect (Huber, Payne, & Puto, 1982) and the compromise effect (Simonson, 1989), which have the form of Ms. X’s preference reversal and both violate the principle of regularity that suggests the preference for Option A should not increase when its choice set is expanded by adding more irrelevant options to it. For the attraction effect, the irrelevant Option D is a decoy (an inferior or dominated option), similar but of less value than A, which creates a bias in favor of A. For the compromise situation, Option C is of approximately equal value to A and B, but it is placed in the middle within the two-dimensional attribute space, making it a compromise. A third and perhaps less puzzling choice reversal is the similarity effect (Tversky, 1972), which violates the independence from irrelevant alternatives principle. Here, the introduction of a new option, S, very similar to B (and of equal value), shifts the relative choice between A and B in favor of the dissimilar option, A. More recently, a new type of reversal effect, the phantom decoy, has been observed (Choplin & Hummel, 2005; Dhar & Glazer, 1996; Pettibone & Wedell, 2000, 2007; Pratkanis & Farquhar, 1992), in which the introduction of an unavailable but dominant option (P in Figure 1) biases the decision toward the similar dominated option (A). Phantom decoy effects raise an additional challenge to the theory of choice (Pettibone & Wedell, 2007).

Such paradoxical preference reversals are, not surprisingly, ruled out by many theories of choice. In particular, they are ruled out by any theory of choice that separately assigns some attrac-

---

Konstantinos Tsetsos, Department of Cognitive, Perceptual and Brain Sciences, University College London, London, England; Marius Usher, Department of Psychology, Tel Aviv University, Tel Aviv, Israel, and Department of Psychology, Birkbeck College, London, England; Nick Chater, Department of Cognitive, Perceptual and Brain Sciences and ESRC Centre for Economic Learning and Social Evolution (ELSE), University College London, London, England.

A substantial part of this work was included in the master’s by research dissertation of Konstantinos Tsetsos, which took place at Birkbeck College, University of London, and was supervised by Marius Usher. Nick Chater was supported by a major research fellowship from the Leverhulme Trust.

Correspondence concerning this article should be addressed to Konstantinos Tsetsos, Department of Psychology, University College London, London WC1H 0AP, United Kingdom. E-mail: k.tsetsos@ucl.ac.uk
Mechanisms for Reversal Effects

Before plunging into details concerning specific models, it is worth considering, in general terms, how a third option might influence the choice between two existing options. There are three broad classes of mechanism based on (a) attentional switching to different choice aspects; (b) relational, rather than independent, evaluation of properties and loss aversion; and (c) value shifts or contrast effects, mediated by lateral inhibition. We consider these briefly in turn.

Attention to Choice Aspects and Temporal Correlations

The similarity effect follows immediately, and fairly uncontroversially, from a stochastic criteria shifting mechanism (Roe et al., 2001; Usher & McClelland, 2004; Usher & Zakay, 1993), a mechanism that has some resemblance to the stochastic examination of choice attributes in Tversky’s elimination by aspects (EBA; Tversky, 1972).

Assume that, while struggling to choose between tiramisu and fruit salad, at some moments, Ms. X is swayed by taste (favoring the tiramisu), and at other moments, she is swayed by health (favoring the fruit salad). That is, her criterion for choice (or in the language of the EBA, her attention to the choice aspects) is continually shifting. Suppose that there is a .60 probability that she will choose fruit salad. However, before she can choose, the waiter points out that there is a third option, fruit surprise, which turns out to be almost exactly the same as, and no better or worse than, fruit salad. Ms. X resumes her oscillations between taste and health.

The structure of the article is as follows. The next section, Mechanisms for Reversal Effects, explores the variety of mechanisms that have been proposed to explain preference reversal and clarifies which mechanisms explain which effects. Then, in Two Neurocomputational Approaches, we describe DFT and LCA in relation to the core theoretical mechanisms and consider the similarities and differences between them. In Distance-Dependent Inhibition in DFT, we take up the challenge of specifying an important parametric dependency in DFT (the dependence of lateral inhibition on the similarity among the alternatives), which was left open in the previous DFT account. The next section, Contrasting DFT and LCA, compares predictions of the instantiations of DFT and the LCA model presented here, and in particular, we raise some apparent problems for the DFT approach and evaluate it against empirical data: This involves limitations caused by local inhibition and linearity and the robustness of the correlational mechanism that accounts for the compromise effect. Finally, in the General Discussion, we summarize and draw conclusions for future research. To anticipate, we find that DFT, as presently formulated, is less robust in the way it accounts for the preference reversal effects compared with the LCA, and we point to a number of experimental predictions that distinguish the two models and motivate future experimental studies.
salad and for fruit surprise are positively correlated (they rise and fall together). In this case, the correlation is caused by the switching of attention to different choice attributes, but as we show below, such correlations can be also caused by other mechanisms. The general idea, however, is that when temporal correlations between momentary preference exist, the correlated options split their wins and, hence, lose share relative to the uncorrelated options.

**Relational Evaluation of Options and Loss Aversion**

The impact of relational, rather than independent, evaluation of options or properties is best illustrated by considering the attraction effect. This corresponds to the addition to the menu of a second tiramisu, which is just like tiramisu but marginally inferior in every way (or, more strictly, marginally inferior in at least one way and no better in any other way). Now consider the relative goodness of each option. If one is not sure how to weigh up the different dimensions of desserts, one may feel that fruit salad is roughly as good as tiramisu and that fruit salad is roughly as good as second tiramisu, but however one weighs the dimensions, it is clear that tiramisu is better than second tiramisu. The specific account of why tiramisu is now relatively favored can take various forms. For example, according to reason-based decision making (Pennington & Hastie, 1993; Shafir, Simonson, & Tversky, 1993; Simonson, 1989), people choose by searching for a justification for their choice. The choice of tiramisu may be justified by its clear superiority to second tiramisu (i.e., it is clearly relatively better, even if one is not sure how much one likes either option, in absolute terms), but fruit salad has no clear justification, being difficult to compare with either alternative option.

Alternatively, both the attraction and the compromise effects could be accounted for, without appealing to a justification process, by assuming that values are computed via pairwise comparisons. For example, we might assume that each option is compared with each other option and that the differences, advantages, or disadvantages (on each dimension, separately) are transformed into utilities via a value function (Tversky & Kahneman, 1991; Tversky & Simonson, 1993) characterized by loss aversion (a steeper slope in the domain of losses than in that of gains, so that losses loom larger than gains).

Consider first the attraction effect (Options A, B, and A’—an inferior decoy of A). The decoy option, A’, now confers to A a clear advantage on both dimensions and thus a net advantage overall. In contrast, A’ confers to B an advantage on one dimension and an almost equal disadvantage on the other. Since the value function makes disadvantages loom larger, the overall value contributed by A’ to B is negative. Exactly the same logic explains the compromise effect. Here, the compromise is the only option that has no large disadvantages being conferred on it from comparisons with other (extreme and with large disadvantages) options in the choice set (Tversky & Simonson, 1993).

**Inhibition as Contrast Enhancement Between Similar Options**

An alternative way to explain the attraction effect is a type of local contrast enhancement, as observed in visual perception (e.g., a circle appears larger when surrounded by smaller circles; Mas-
allowed to go negative. The second nonlinearity is carried over from prospect theory, in the form of an asymmetric value function with loss aversion (losses weighted higher than gains), which is taken by LCA as a primitive. Unlike the LCA, which maintains most of the aspects of Tversky’s theories, DFT does not assume loss aversion as a primitive but rather derives it as an emergent property. To do so, it assumes that the inhibition between the choice alternatives is an increasing function of their similarity in the attribute space. Despite the central role of the decreasing inhibition-distance function, no explicit function was used in DFT (and as we show later, the choice of the inhibition function turns out to be important), aside from the special case of the step function. We start with a brief description of DFT and LCA models (the text focuses on main principles; a detailed description is presented in the appendices), and then, we examine the choice patterns that DFT generates under various inhibition-distance functions. After characterizing the inhibition mechanism in DFT, we proceed with a set of comparisons between the two models and discuss some difficulties in the current DFT formulation for both the attraction and the compromise effects.

Reviewing DFT and LCA

DFT and LCA are both instantiated in four-layered connectionist networks as illustrated in Figure 2. The first layer corresponds to the choice attributes (two attributes are illustrated here). In both models, it is assumed that the attention of the decision maker switches stochastically across dimensions (D1, D2), according to a Bernoulli process3; hence, at any time step, only one of the attributes is active. The two-dimensional characterization of each alternative on the D1–D2 space (see Figure 3) is given by the connectivity between the first and the second layers (i.e., a 2×3 matrix). Each node in the second layer corresponds to the integrated attribute values of each choice alternative (see Appendix A, Equation A1).

Figure 2. Illustration of decision field theory and leaky competing accumulators models in neural networks; circle arrow heads correspond to inhibition. a: Connectionist network for decision field theory. b: Connectionist network for leaky competing accumulators.

The two models differ slightly on the intermediate computations performed in the third layer and on the way in which the preferences are integrated in the fourth layer. In DFT, the third layer computes contrasts between each option and the other alternatives (also mentioned as valences) as the difference between the value of the option and the mean value of the other options, with respect to the active dimension (taken from the second layer; see Appendix A, Equation A2). In LCA, the third layer computes advantages and disadvantages between all pairs of options, which are transformed by a nonlinear, asymmetric (loss-averse) value function (see Appendix A, Equation A4). Finally, in both models, the fourth layer integrates the contrasted differences (valences or sum of advantages/disadvantages in DFT and LCA, respectively) as preferences across time.

The integration of preference for each option is imperfect (leaky) and subject to competition with the preferences of the other options (see Appendix A, Equations A4 and A5, for DFT and LCA, respectively). The leaky integration of preferences and the competitive interactions between the options are implemented in a connectivity matrix, whose diagonal term corresponds to a self-connectivity coefficient (or the leak parameter chosen as .94 in the simulations presented here, unless stated otherwise) and whose off-diagonal elements correspond to inhibitory connections. While, in LCA, all the off-diagonal elements are constant (global inhibition), in DFT, their magnitude depends on the distance between the alternatives (in the two-dimensional attribute space). Finally, as mentioned above, DFT is linear, and thus, preference states can take both positive and negative values, as opposed to LCA, where negative activations at the fourth layer are truncated to zero. In the original DFT model for preference reversal (Roe et al., 2001), the connectivity matrix, s, is such that its eigenvalues are smaller than one, preventing unstable dynamics that result in unbounded activation levels. This poses a restriction on the class of inhibition functions (the off-diagonal terms). This restriction can be relaxed by using an additional mechanism that prevents unbounded activation (J. Busemeyer, personal communication, November 4, 2009).

While the two models explain identically the similarity effect, their explanations for the attraction and compromise effects are very different. In DFT, it is the contrast enhancement mediated by local inhibition that accounts for the attraction effect; the value of the dominating option, A, is enhanced by the similar decoy. In particular, the similarity between nearby alternatives (A and D in Figure 3b) results in their being coupled by strong local inhibition. As Option D is inferior to both A and B, it has negative valence. Therefore, Option D boosts the preference of Option A by passing its negative activation value through a negative connection (we call this activation by negated inhibition). The function that specifies the local inhibition relates the psychological distance (i.e., similarity) of the options and the degree they compete by lateral inhibition.

The compromise effect is also accounted for by DFT due to the distance-dependent inhibition; however, the key mechanism is correlation, not contrast enhancement. In this case (see Figures 3c and 3d), the extremes (A and B) and the compromise (C) interact

---

3 More complex models of the shifting of the attention across dimensions are possible, for example, models with the Markov property (Diederich, 1997).
via strong inhibitory links, whereas the extremes, A and B, are too
distant from each other to compete. As the extremes do not inhibit
each other, although they inhibit the compromise option, their
momentary preference becomes decorrelated from the compromise
but correlated with each other. Thus, the correlated extremes split
their wins, making the compromise option stand out and take a
larger share of choices (see Roe et al., 2001, for details).4 The DFT
model has a different way to account for the compromise effect,
when its $s$ matrix has eigenvalues larger than one that result in
unstable dynamics. For example, the situation may be such that
adding Option B to the pair A and C makes the $s$ matrix unstable
(C now being linked by inhibition to two options instead of one).
In that situation, the C activation will go to $\pm$ infinity, depending
on noise; thus, for options of equal valence, C will win half of the
time, while the extremes will share the other half (J. Busemeyer,
personal communication, November 4, 2009).

Unlike in DFT, the LCA account of the attraction and the
compromise effects is similar to the context-dependent advantage
model (Tversky & Simonson, 1993) and does not require a
distance-dependent inhibitory mechanism. Instead, it follows the
principles suggested by Tversky and Simonson (1993), according
to which the value for each option is evaluated in relation to all
other options in the choice set (so far, this is not fundamentally
different from DFT) via a nonlinear loss-aversion value function.
In particular, for the attraction effect (see Figure 3b), when Option
D is introduced, Option B is penalized more by having two large
disadvantages (relative to A and D, when dimension of economy
is attended to) relative to A (which has one large disadvantage
only). The same principle helps the LCA account for the compro-
mise effect (see Figure 3c); the extreme options (A and B) have
one large and one small disadvantage each, whereas the compro-
mise option has two small disadvantages. Due to the asymmetry of
the value function, large disadvantages are penalized more, favor-
ing that way the compromise option. A summary of the accounts
that each model gives for each effect is given in Table 1.

The inhibition function, which is crucial to the explanatory
power of DFT, is the first topic of this investigation. As shown by
Roe et al., 2001, it is possible to find inhibition values that capture
all three effects.5 However, since Roe et al. examined only ordinal
distance relations between alternatives (similar/dissimilar), an ex-
licit functional specification for the distance function is needed to
make parametric predictions for DFT. One such function, consis-

Figure 3. The choice sets corresponding to the three effects and annotation of the distances between the options
that determine the inhibition values in decision field theory. a: The similarity effect. b: The attraction effect. c:
The compromise effect. d: Explicit inhibition values that can account for the three effects simultaneously. In
Panel d, on the y-axis: $H$ = high, $L$ = low; on the x-axis: $S$ = small, $M$ = medium, $L$ = large.

4 For both the compromise and similarity effects, DFT gives a correla-
tional account. However, while, in the similarity effect case, the temporal
correlations occur between the similar options as a result of the attentional
switching, in the compromise effect case, the correlations occur between
the two extreme (dissimilar) options as a result of the distance-dependent
inhibition.

5 Figure 14 in Roe et al. (2001) shows that the model can account for the
three effects for a variety of noise/inhibition parameters.
tent with (though not suggested by) the DFT model, could be a step function, as depicted in Figure 4b (red curve): Inhibition is high within a range and, outside it, is virtually zero. Since, in psychological theories of similarity, a step function is unusual (Nosofsky, 1986; Shepard, 1987), we explore here other inhibition functions that can account simultaneously for the three effects.

### Distance-Dependent Inhibition in DFT

#### Linear, Exponential, and Gaussian Functions

The aim of this section is to explore the distance-dependent inhibition function that allows DFT to explain the three phenomena simultaneously. Before we start, we note that these effects were so far obtained in different studies, so until a study reports all three effects with the same materials, procedures, and subjects, there is the possibility that more freedom exists if parameters (e.g., noise) can be modified for various decoy effects. We mainly consider here the same-parameter case, but we also discuss some other possibilities. We started with the simplest type of decreasing functions of distance, which are piecewise linear. Next, motivated by well-known theories of similarity (Nosofsky, 1986; Shepard, 1987), we focused on exponential and Gaussian functions of inhibition. The results were obtained using Monte Carlo simulations and keeping the noise parameter constant to .2 and the leak parameter to $\lambda = .94$. None of the linear and exponential functions was able to capture the three phenomena simultaneously (the details are omitted here, but see Tsetsos, 2008, for details). The Gaussian inhibition functions we tried are illustrated in Figure 4a. We crossed the starting point of the inhibition (three values) with different slopes. The results for the three effects are summarized in Table 2, suggesting that the Gaussian functions also fail to account for the three reversal effects simultaneously.

We believe that the reason DFT cannot capture the three phenomena with smoothly decaying inhibition functions, such as the Gaussian functions, is the following. The similarity effect under the DFT framework is maximally obtained for global inhibition, but it still can be obtained when the inhibition at small distances (B vs. S) and at large distances (A vs. B) does not differ a lot. However, the compromise effect requires a large difference in the inhibition between intermediate (A vs. C) and large distances (A vs. B). To satisfy these conditions together, the distance function needs to decay slowly or not at all over intermediate distances but with a much higher slope at large distances (see also Figure 3d for such an extreme case); Gaussian functions do not decay slowly over intermediate distances and significantly faster enough at large distances.

In our initial explorations, we found a sigmoid (logistic) function that can satisfy all three reversal effects (Tsetsos, 2008; see also Figure 4b, red line). Another inhibition function that can satisfy the three effects together has recently been proposed (J. Busemeyer, personal communication, November 4, 2009). This function, which is a Gaussian of the distance-square, has a sharper decay than a normal Gaussian (black line in Figure 4b). Below, we refer to this more abrupt distance function as the Gaussian-distance-square, and we use it along with our sigmoid function (when the two functions provide

### Table 1

A Summary of DFT and LCA Accounts of the Preference Reversal Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>DFT</th>
<th>LCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity</td>
<td>Attentional switching across dimensions</td>
<td>Attentional switching across dimensions</td>
</tr>
<tr>
<td>Attraction</td>
<td>Excitation by negated inhibition</td>
<td>Loss aversion in value function</td>
</tr>
<tr>
<td>Compromise</td>
<td>Correlations due to local inhibition</td>
<td>Loss aversion in value function</td>
</tr>
</tbody>
</table>

*Note. DFT = decision field theory; LCA = leaky competing accumulators.*

---

6 In addition, it is assumed here that the psychological distance between two options increases more rapidly along the line of dominance and less rapidly along the line of indifference. Intuitively, this new metric of distance suggests that options with equal additive utilities compete more strongly, while inferior options appear distant and do not interact with superior options. This concept is expressed by transforming the conventional distance between two options into the sum of the squares of the two new dimensions of indifference and dominance (see Appendix B) and then applying a larger weight to the dimension of dominance.
distinctive predictions) in all the following DFT simulations. First, though, note that the mechanism for obtaining the compromise effect with the Gaussian-distance-square inhibition of Figure 4b is based on the transformation of the s matrix from a stable to an unstable one, when Option B is added to the pair A, C (note that the off-diagonal inhibition magnitude is higher for the Gaussian-distance-square compared with the sigmoid).

### Contrasting DFT and LCA

In this section, we explore parametrically how choices depend on the locations of the choice alternatives in the attribute space. Specifically, two options (i.e., A and B) remain constant, while the third option (i.e., C) moves across the two-dimensional space with an increment of .05 at each step. We consider only results at or above the diagonal between A and B, as, above the diagonal, Option C is always chosen. For the DFT model, we use the Gaussian-distance-square function defined on the indifference/dominance directions (see Appendix B, Equation B2). The parameters that were found to optimize DFT in predicting correctly the reversal effects were $\sigma = .05$ (additive noise, see also Appendix A, Equation A4), $\varphi_1 = .022$, $\varphi_2 = .05$, $b = 12$ (J. Busemeyer, personal communication, November 4, 2009). For the sigmoid function we used noise $\sigma = .2$ and inhibition $= .042$ as the starting point of inhibition, while the function started to decay after a distance $d = 2.15$ and with a slope equal to $s = 20$. For the LCA model, in preliminary investigations (Tsetsos, 2008), we found predictions to be robust to the value of the global inhibition and to the value function used, as long as it is asymmetric, such that disadvantages are weighted more highly than advantages (Kahneman & Tversky, 1979). Note that the LCA with asymmetric value function results in attraction and compromise effects co-occurring.² For brevity, we present here the results obtained using the value function from prospect theory (Kahneman & Tversky, 1979):

$$V(x) = x^{2.5}$$, $x < 0$.

A representative set of results (for $I_0 = 2$ and $\lambda = .94$) is presented in Figure 5c, along with the predictions of the DFT model with the Gaussian distance-dependent inhibition (Figure 5a) and with the sigmoid function (Figure 5b). The figure illustrates the magnitude of the attraction and similarity effects with respect to Option A, as the difference between the probability of choosing A and the probability of choosing B, for different locations of Option C in the two-dimensional lattice. We use a gray scale, where brighter points correspond to a stronger enhancement of the preference of A by the introduction of C. For both models, we can see the similarity effect illustrated as a thin white line close to Option B (1, 3) and adjacent to the diagonal (i.e., the introduction of Option C similar to—neither dominating nor dominated by—Option B results in boosting the preference for the dissimilar Option A). The predictions for the attraction effect diverge, however. For the LCA model (see Figure 5c), the attraction effect is present in the triangular white area close to Option A. The magnitude of the effect gradually decreases as the distance between the decay (Option C) and the target (Option A) increases. This prediction is a consequence of the asymmetry in the value function, which renders the relative disadvantages of the competitor (Option B), which determine the magnitude of the attraction effect, dependent upon the position of the decay (Option C). On the other hand, DFT gives a more dichotomous prediction regarding the magnitude and the location of the attraction effect for both the distance functions we used. As Figures 5a and 5b illustrate, there are areas in which the attraction effect occurs and areas in which it does not (the precise range depends on the parameter of the distance function in the dominance direction for the distance-square function). More importantly, the DFT predicts that the magnitude of the effect is relatively flat within the area where it takes place. This discontinuity directly stems from the relatively abrupt distance inhibition functions. Since there are no empirical findings that clearly relate the magnitude of the attraction effect to the distance between the target and the decay, we do not see these

² It is possible that some subjects show attraction without compromise effects. The LCA framework is able to account for this with a symmetric value function (Bogacz, Usher, Zhang, & McClelland, 2006).
Avoiding Dominance Reversals in DFT

The attraction effect in DFT is a type of contrast effect in which the decoy enhances the dominating option with which it is contrasted. While this works well in the attraction situation, this mechanism has the danger of causing dominance reversals for options that are in a strict domination order, as illustrated in Figure 6a (C dominates B, and B dominates A). Such reversals may occur (depending on the magnitude of the inhibition), when the distance between the options is such that A and B inhibit each other while C is more distant and outside the inhibition range of the two dominated options.

In the simulation result in Figure 6b we used an inhibition value of .049 (between A and B), consistent with the localized distance functions in Figure 4b that allowed us to reproduce all three reversal effects. As can be seen, although the activations are bounded (all eigenvalues of the s matrix are smaller than 1), after approximately 200 time steps, the dominated Option B emerges as the choice winner (Figure 6c shows a single-trial trajectory for DFT). This result can be confirmed analytically by the steady state of the system, which is given by

$$P = (I - S)^{-1} \times V,$$

where $V$ is the valence matrix (see Appendix A, Equation A2) and $S$ is the connectivity matrix derived from the distance-square function of inhibition (see Figure 4b, solid black curve). For this example, $A = (0, 0)$, $B = (2, 2)$, and $C = (1, 1)$, and the steady state is $P(A) = -72$, $P(B) = 62$, $P(C) = 18$ (see Appendix C for derivation of this and for the mapping of inhibition levels that produce such dominance reversals). Intuitively, this prediction results from the fact that the superior option, C, does not benefit from the boosting by negated inhibition from any option since it does not interact with either A or B. On the other hand, the inferior decoy, A, which has a negative valence, confers excitation on B. On the contrary, as illustrated in Figure 6d, the LCA gives the correct prediction since, due to the nonlinearity in the preference function.

Figure 5. Illustration of the attraction and similarity effects as the boost that A gets relative to B by the introduction of C (i.e., $P(A|A, B, C) - P(B|A, B, C)$) in various places of the two-dimensional lattice. a: Predictions for decision field theory (DFT), distance-square inhibition function. b: Predictions for DFT, sigmoidal inhibition function. c: Predictions for leaky competing accumulators.

Figure 6. a: A choice scenario where Option C has the highest additive utility. b: Probability of choice for the three options in decision field theory (DFT); after approximately 200 time steps, the inferior Option B outplays C due to the sharp boundaries of the inhibition (see Figure 4b, black line). c: Single-trial trajectory for DFT. d: Predictions for the leaky competing accumulators model.
states, uninformative options are deactivated (stuck at zero) at early stages of the decision process.

There are a number of ways in which the DFT can address these problems. First, one can decrease the inhibition magnitude. In Appendix C, we show that the critical magnitude that triggers dominance reversal is lower than the one that triggers instability and that it depends on the leak (\(\lambda\)) and on the distances between the three options. As shown in Appendix C, one can prevent dominance reversal by reducing inhibition below the critical value of .043 while still maintaining similarity and compromise effects (see Table C1 in Appendix C for exact parameters). A second way to prevent dominance reversals is by imposing a negative boundary on activations such that activations cannot go lower than this value. Third, one can limit the effective time that the decision makers are engaged in the deliberation process via an attentional slowdown process (J. Busemeyer, personal communication, March 20, 2010). We thus do not see these problems as critical (but we think that they motivate further model development in the DFT framework), and we now turn to a set of comparisons between predictions of DFT and LCA for the decoy effects.

**Distance Dependency for the Compromise Effect**

The local inhibition mechanism also has consequences for the range of the compromise effect, as a function of the separation between the extreme options. To examine this, we contrasted the DFT and LCA choice patterns for triples A, B, C, in which the compromise option, C (2, 2), is constant while the distance \(x\) between the extremes varies in a symmetric fashion (see Figure 7a). In Figure 7b, we report the magnitude of the compromise effect, as a function of the distance between the extremes, for both DFT and LCA (positive values correspond to an advantage of the compromise option, negative values to a disadvantage).

We observe (see Figure 7b, black and red lines) that the compromise effect in DFT occurs only within a particular distance range, which is directly determined by the distance threshold of the inhibition function; the effect occurs only when the compromise is within the inhibition range from the extremes, and it disappears when the extremes are in inhibition range with each other. This holds for both the sigmoid (red line) and distance-square (black line) inhibition functions as both are localized. In the latter, case however, as the effect is mediated by the unstable s matrix, we find that the effect has little robustness with regard to small changes in the valence of the compromise option, C. For example, it was enough to change C from (2, 2) to (1.99, 1.99) to reverse the compromise effect (C now wins only 1%). Finally, if we relax the assumption that the similarity and compromise effects have to be predicted by the same parameter set and allow for other parameters such as noise to vary for each effect, we can have less abrupt predictions (cyan line; for higher noise, \(\sigma = .7\) applied for the compromise effect) for a normal Gaussian distance-dependent inhibition function (see Figure 4b, blue dashed line). Again, however, the compromise effect becomes small at large distances (i.e., when the compromise becomes outside the range of the local inhibition of the extreme options).

By contrast, the LCA model predicts (see Figure 7b, black line) a robust compromise effect that is continuous with the separation between the extremes, and its magnitude increases proportionally to the separation, subject to saturation. This reflects the properties of the asymmetric value function; the slope in the domain of losses increases with distance, penalizing the extreme options. Future empirical studies should assess the robustness and the magnitude of the compromise effect as a function of the distance between the two extremes.

**The Correlational Nature of the Compromise Effect in DFT**

According to the original DFT model (Roe et al., 2001), the compromise effect occurs because the preferences of the extremes are correlated in time. A different way for the DFT to explain the compromise effect is related to the change of the s matrix into an unstable one upon the addition of a new alternative. These accounts of the nature of the effect are very different from the one offered by the LCA, which follows Tversky and Simonson’s (1993) proposal that it is a result of the fact that options are compared to each other and that large disadvantages are penalized. Here, we compare these two types of account in two ways. First, we argue that the correlational account lacks robustness when
additional options are added. Second, we consider experimental results that directly address the correlational nature of the effect.

**Robustness of the Correlational Mechanism**

The correlational nature of the compromise effect is based on the fact that the extremes are in inhibition with the compromise option but not with each other and that therefore they become decorrelated from the compromise and correlated with each other. When DFT dynamics are stable, the correlational mechanism depends both on the strength of the inhibition between the compromise and the extremes and on the level of noise. This mechanism is not robust to situations in which additional options are introduced. Consider the case with five alternatives, illustrated in Figure 8a, in which we have four (instead of two) extreme alternatives. As the two similar extreme pairs (A-D and B-E) are mutually inhibitory, the decorrelation from the compromise is unlikely to be enough to make the extremes more strongly correlated than the compromise. To show this, we ran DFT simulations corresponding to this five-alternative choice set. Here, we relied on our sigmoid inhibition function, and we lowered the leak to .93 to maintain stability for the five-alternative choice scenario.

We observe (see Figure 8b) that the compromise effect now disappears. When we set the inhibitory connections within the similar pairs (A-D and B-E) to zero, the compromise effect is restored (see Figure 8c), demonstrating that it is the local inhibition that is the factor responsible for its disappearing. Future experimental work is required to examine the robustness of the compromise effect in a five-alternative choice scenario.

**Testing for Correlations Between Preferences**

If the compromise effect arises from the temporal correlation of the extremes, it should be possible, in principle, to detect a signature of this correlation. One way to investigate this was recently explored experimentally (Usher, Elhalal, & McClelland, 2008). In this study, participants were presented with a three-choice compromise choice set, and in some cases, following the participants’ choice of an extreme option, this option was announced to be unavailable, and a speeded second choice was requested for one of the remaining two options (see Figure 9a). If the two extremes indeed become correlated, one may predict that, at the moment when one of them reaches a response criterion, the other extreme is also high in its activation and therefore is more likely to be selected in a speeded choice. Computer simulations of the DFT (Tsutsos, 2008) confirmed the statement that, unlike in LCA, if the preference value of an extreme option is inhibited (so that it is eliminated and rendered unavailable from the choice competition) after it reaches the response criterion, the other extreme is more likely to be selected, especially at short time intervals.

The experimental results in Figure 9b show that after the choice of an unavailable extreme, the participants had an overwhelming tendency to choose the compromise, rather than the other extreme (Usher et al., 2008). Furthermore, the selection times were longer when participants chose the other extreme than when they chose the compromise (see Footnote 9; also see discussion in Usher et al., 2008).

Our results do not confirm the correlational account of the compromise effect assumed by the DFT model. However, within the DFT framework, an alternative explanation for this effect has been suggested (Busemeyer & Johnson, 2004). According to this account, availability can be seen as a third choice attribute, which makes an unavailable option less desirable but allows it to compete for selection. According to this, the unavailable option bears negative valence due to its low attribute value in the availability dimension and boosts the compromise due to their mutually inhibitory connections (negated inhibition). Such a mechanism could therefore provide an alternative account for the data from this experiment.

This availability mechanism can be tested as follows. Consider a choice set with three options, as in the attraction case. During the deliberation, we announce that the decoy is unavailable. According to the availability assumption, this will make the valence of the decoy option even more negative, and thus, the boost it should give to the dominant option should be further enhanced. On the other hand, if unavailable options are simply eliminated from the choice set, we would expect that the attraction effect will diminish toward the baseline for a binary choice. To test this, we presented to 30 participants (students at University College London, London, England) three choice problems, all of which involved the same two alternatives, A and B, which created a trade-off between two choice attributes. The first problem was a binary choice between A and B. The second problem was a trinary choice in which a decoy dominated by A and similar to it was added. The third problem was identical to the second, except that after 15 s of deliberation, the decoy was announced to the participants as unavailable. The participants were divided into three groups, each of which received a different permutation of three problems with three types of material (visit to clinic, choice of flowers, and candidates for a master of science scholarship). The problems were presented as a PowerPoint presentation, and the response was solicited after 30 s (no earlier responses were accepted). Thus, for decisions with unavailable products, this information was given halfway within the decision time.

The results, reported in Figure 10, are clear and do not fit with this alternative explanation of the availability being an extra choice dimension. The decoy induced a strong attraction effect in favor of the dominating option, $\chi^2(1, N = 30) = 6.67, p < .01$, between trinary and binary. When the decoy was announced as unavailable during the deliberation, its impact disappeared, and the choice between Options A and B reversed very close to baseline, $\chi^2(1, N = 30) = .07, p > .78$, between trinary-unavailable and binary.

---

8 Note that the distance-square inhibition function results in unstable dynamics even in the three-alternative compromise effect case.

9 One caveat to this prediction is that, following the announcement of the unavailability of the preferred choice, the participant will restart the choice from scratch, in which case the advantage of the correlated alternative becomes immaterial. Such a restart, however, is expected to lead to longer choice latencies, leading to a second prediction: The choice latencies of the second response (following the unavailability input) should be faster when the other extreme is chosen than when the compromise is chosen (as, in the latter case, a restart is more likely). The LCA model makes the opposite prediction. See Usher et al. (2008) for results and discussion.

10 The unstable explanation is also contradicted by these data. According to this, the extremes go together to either positive or negative activations.
The results of this experiment are also different from the ones regularly obtained with phantom decoys (Choplin & Hummel, 2005; Dhar & Glazer, 1996; Pettibone & Wedell, 2000, 2007; Pratkanis & Farquhar, 1992), since, unlike in the latter, in our experiment the effect of the unavailable decoy dissipated, resulting in a binary choice baseline preference pattern. Note, however, that unlike our decoy, which is an inferior (dominated one) the decoys used in phantom decoy studies are superior to the target. As discussed in detail by Pettibone and Wedell (2007), relational valuation models with loss aversion such as LCA and the context-dependent advantage model can account for the phantom decoy effect by assuming that people use the superior unavailable decoy as a reference point. The DFT could also account for phantom decoys if the valence of the superior decoy is negative (due to its unavailability); however, it has difficulties in explaining how the magnitude of the effect depends on its distance from the target (Pettibone & Wedell, 2007).

**General Discussion**

DFT and the LCA are computational models of multiattribute decision making, which can account for contextual reversal effects. The two models share many properties and use similar connectionist frameworks, but they differ in the way they account for the attraction and compromise effects. While the LCA follows the more traditional account offered by Tversky and Simonson (1993), in which the effects arise from the asymmetry of the value function and the fact that options are compared with each other, DFT does not assume asymmetric loss-aversion value functions. Instead it derives the attraction and compromise effects from the emergent properties of the local inhibition within a linear network. The attraction effect is viewed as a contrast effect, which results from the fact that the decoy boosts the preference of the similar dominating alternative by the mechanism of activation by negated inhibition. The compromise effect is the outcome of an emergent correlation between the extremes, which share their wins in the choice, favoring the compromise option.

To compare the models’ predictions, we examined the family of distance-dependent inhibition that enables the DFT to account for the three reversal effects with the same model parameters. We found that this inhibition function needs to have a relatively abrupt boundary (see Figure 4b, red line), to reproduce simultaneously the similarity, the attraction, and the compromise effects. Another localized function (see Figure 4b, black line) can account for the effect under the unstable matrix scenario (but this has little robustness to changes in the valence of the compromise, as we showed in the Distance Dependency for the Compromise Effect section).

Figure 8. a: A five-alternative choice problem similar to the compromise effect case; the all-average option (C) is expected to win. b: Probability of choice for the five options in decision field theory (DFT); the compromise option is not chosen. c: Probability of choice in DFT when the inhibitory links in the pairs of the extreme options (A, D and B, E) are removed. d: Probability of choice for the five options in the leaky competing accumulators model; the compromise option is chosen.
With these inhibition functions, we compared DFT and LCA’s predictions for multiattribute decisions.

Our simulations showed that, as a result of the local inhibition boundary, the range of the attribute space in which DFT produces reversal effects also has relatively sharp boundaries, which stand in contrast to the more continuous effects obtained in the LCA model, and which could be subject to future experimental investigations. We also found that the predictions of DFT are less robust to the introduction of new options in the choice set (see Figure 8a) and that the inhibition level needs to be properly restricted to prevent dominance reversal (see Figure 6). This restriction results in a smaller parameter space for noise and inhibition than the one that was possible in Roe et al. (2001). We expect, however, that additional mechanisms could be used to allow DFT more robustness in dealing with such problems. For example, a mechanism may be required to restrict activations and depart from linearity, as some of the problems are produced by the local inhibition combined with the linear dynamics, which allows options with low valence to boost their nearby options and that way to distort the choice process. This stands in contrast to the LCA, which uses its nonlinearity to eliminate inferior option from the choice.

Finally, we examined whether the compromise effect should be explained in terms of correlations, which is probably the most original mechanism in the DFT account of multiattribute decision making. To examine the correlational prediction, we introduced a choice situation in which decision makers are presented with a choice between three alternatives that form a compromise situation and, following the choice of an extreme option, that option is announced as unavailable and another speeded choice is requested. Using this experimental design, Usher et al. (2008) found that following the choice of an extreme option, the overwhelming fraction of speeded choices goes to the compromise option, rather than to the other extreme. In one version of DFT (Busemeyer & Johnson, 2004), such a result can be accounted for by assuming that unavailability is a third attribute, which does not eliminate an option from the choice process but only reduces its valence, making it less attractive. Under such a mechanism, the unavailable extreme plays the role of the decoy, which activates the compromise via activation by negated inhibition. Note that for this to happen, the unavailable extreme needs to continue to interact with the available options (rather than being dropped from the decision) and have a negative valence. To test this, we carried out an experiment, which compared the attraction effect in a normal situation to that in a situation in which the decoy option is set unavailable after 15 s of deliberation. We found, that, converse to the prediction that the unavailability of the decoy reduces its valence, enhancing the attraction effect, this effect is reduced toward the baseline of binary choice. This suggests that unavail-

![Figure 9](image_url)

**Figure 9.** a: When one extreme option (A) is chosen, it is set unavailable, and the remaining options (B and C) compete until a new decision is made. b: The experimental results show that in the second choice between the compromise (C) and the available extreme (B), participants dramatically preferred the compromise option. Error bars indicate 95% confidence intervals of the mean.

![Figure 10](image_url)

**Figure 10.** Experimental results testing the role of unavailable options in the deliberation process. The choice pattern of the binary case is reversed in the presence of a decoy option (i.e., the attraction effect). However, when the decoy option is set unavailable after 15 s of deliberation, then the preference falls back to the baseline of the binary case. Error bars indicate 95% confidence intervals of the mean.

---

11 According to this mechanism and counter to introspection, when a person feels ready to choose an extreme option out of a compromise set, the preference for the other extreme is also quite high and is stronger than the preference for the compromise.

12 It could be argued that the effect of inhibition might dissipate with time and disappear after 15 s of deliberation following announcing the dominated option unavailable. This, however, would mean that the unavailable decoy is either chosen or simply eliminated. In the latter case, it could not affect decisions.
ability should not be viewed as a third choice dimension (making unavailable options slightly less desirable), but rather that it maintains their desirability while eliminating them from the decision process.

While the results of these comparisons present challenges to DFT, it remains possible that a revised specification may meet these challenges successfully. There are several ways in which DFT may potentially be strengthened further. First, note that the noise parameter, if allowed to vary across different choice sets, can play an important role in allowing a more gradual inhibitory distance function to account for the three effects simultaneously (J. Busemeyer, personal communication, March 20, 2010). Equally, note that the three effects have not all been obtained in a single experiment. It is thus conceivable that there is something in the procedure used by Tversky (1972) that corresponds to a lower degree of noise (or attention to irrelevant dimensions) than the experimental procedure used by Huber et al. (1982) and by Simonson (1989). Future studies are thus needed to test if the three effects can be replicated within the same experimental design. If, indeed, the similarity effect requires less noise in its procedure, this may allow more space for a gradual inhibition function to account for reversal effects in the DFT model.

Another factor that could help strengthen DFT is the introduction of some nonlinearity in its dynamics. The nonlinearity is in fact one of the differences between the DFT and the LCA. The linear dynamics of DFT are attractive, from a theoretical standpoint, because predictions can be obtained using linear algebra rather than time-consuming simulations. In LCA, by contrast, the nonlinearity is motivated in terms of neurocomputational principles (activation is seen as corresponding to firing rates and thus cannot go negative) and enhancing computational efficiency. For example, due to its zero-threshold nonlinearity, the LCA can naturally eliminate inferior options from the decision process, preventing them from introducing noise that may distort the decision preference (see Bogacz et al., 2007, for a discussion of the advantage of nonlinearity in perceptual choice). This is closely related to the issue of biological plausibility; in their response to Usher and McClelland (2004), Busemeyer, Townsend, Diederich, and Barkan (2005) suggested that the activation by negated inhibition in DFT could be understood biologically in relation to a disinhibition process. Note, however, that the phenomenon of disinhibition should be bounded. Consider an inhibited target neural unit (T in Figure 11a) that fires below the baseline (i.e., negative in DFT) on account of the inhibition it receives (from Unit A in Figure 11a). If the inhibiting neuron is suppressed, then the firing rate of the previously inhibited unit (T) potentially exceeds the baseline rate (see Figure 11b). However, the firing rate of the disinhibited unit cannot be higher than its firing rate when it is no longer linked with an inhibitory connection with the inhibiting neuron. In other words, the boost that the previously inhibited neuron gets from disinhibition cannot be infinitely large but must be bounded from above (say, via its excitatory input, B, in Figure 11b). To satisfy this biological constraint and to also prevent instabilities, DFT may need to introduce either an upper or a lower bound on the value of activations, thus departing from linearity.

Such a mechanism was indeed discussed by Roe et al. (2001, p. 385) in considering how strategy switching can be implemented in DFT. Future implementations of such nonlinearity in DFT need to be tested for the various choice situations. Such a model would face a dilemma: to assume whether options that reach the low boundary are fully eliminated from the choice or still influence the decision by competing with the alternatives that are within their inhibition range. An alternative account could be to limit the effective time of decision making by reducing the gain with which the target equations proceed as the decision unfolds (J. Busemeyer, personal communication, March 20, 2010).

One of the distinctive features of DFT is the fact that it produces loss aversion as an emergent property without assuming any asymmetry or nonlinearity in the value function. In LCA, on the other hand, we have followed Tversky and Kahneman (1991) in their assumption that the asymmetry between gains and losses is a primitive, which is hardwired in the neural system. Although, no consensus on the nature of loss aversion has yet been reached, a recent study of brain imaging during risky decision making supports the hypothesis that there is a single system in the brain that encodes subjective value asymmetrically, weighing disadvantages more than advantages (Tom, Fox, Trepel, & Poldrack, 2007). Such findings, however, do not rule out explanations of the value function asymmetry, reflecting environmental contingencies. For example, in decision by sampling (Stewart et al., 2006), loss aversion is attributed to the asymmetry of the real world distributions of gains and losses.
Alternative Models

In this article, we have focused only on DFT and LCA as they are the only two theories that account for the similarity, attraction, and compromise effects simultaneously. Alternative theories have been proposed for multialternative, multiattribute choice, namely, decision by sampling (Stewart et al., 2006) and the ECHO model (Guo & Holyoak, 2002), with the latter accounting for a subset of the reversal effects. Three particular mechanisms stand out from the two models as promising: ranking, grouping, and bidirectional connections in the neural network.

In decision by sampling, no underlying psychoeconomic scales are assumed. Instead, the subjective value of an attribute is its rank in the decision sample, which consists of attribute values both present in the decision context and drawn from memory. Thus, the value of a given option is constructed online using basic cognitive tools such as binary comparisons and frequency accumulation. Drawing from simple psychological principles, decision by sampling accounts for a large set of decision phenomena such as loss aversion, temporal discounting, and overestimation of small probabilities. Being explanatorily robust in several domains, decision by sampling and its mechanisms (ranking and ordinal comparisons) appear to be promising for the case of preference reversal effects. Recently, decision by sampling was integrated with LCA in a dynamical model for decisions under risk (Stewart & Simpson, 2008). This model can also be extended for multiattribute decisions, and its descriptive power in that domain should be the subject of future computational explorations.

The second alternative model, the ECHO model proposed by Guo and Holyoak (2002), has been applied for the similarity and the attraction effects. Its central assumption is that decisions follow a sequential two-stage process. At the first stage, the two similar options are grouped and processed together. The preference states of the similar options from the first stage are carried over as initial activations at the second stage, where all three alternatives are compared together. Thus, the similar, grouped options receive more processing time overall. Note that the mechanism of grouping can be comparable to the step sigmoid inhibitory function in DFT, which involves competition only between the similar options.

Another assumption in the ECHO model is that the preference states of the alternatives are passed backward to the attribute nodes, providing positive feedback. Therefore, it is predicted that during deliberation, the attribute values of the option that is dominating the preference will be enhanced and thus appear to be more important, which has been confirmed experimentally (Holyoak & Simon, 1999). It would be interesting to test what further predictions the LCA and DFT models would yield by changing their connectionist networks from feedforward to bidirectional (Glöckner & Betsch, 2008).

To conclude, we have pointed out some difficulties of the DFT account of multiattribute decision making and suggested a number of ways to test between the DFT and LCA models. We believe that future experimental and computational investigations are needed to develop a solid neurocomputational theory of multiattribute decision making. Such a theory may share assumptions with both DFT and LCA, as well as other decision-making models.

References


### Appendix A

**Dynamics of the DFT and LCA Models**

In this appendix, we show the computations that are performed at each layer of the decision field theory (DFT) and leaky competing accumulators (LCA) networks. At the first layer for both the models, a stochastic mechanism allocates the attention among the attribute units, so that only one attribute (randomly determined) provides input to the second layer at each time step. The second-layer activations are common for the two models and are determined by the following equation:

\[
U_i(t) = \sum_{j=1,2} w_j(t) \times m_j + \varepsilon_i(t).
\]  

(A1)

In the above equation, \(\varepsilon\) is the probability of attending irrelevant dimensions, and \(w_j\) is either 0 or 1 depending on which dimension the attention is focused.

The second-layer activations are passed forward to the third layer. For DFT, the valences are computed as

\[
v_i(t) = U_i(t) - \left( \sum_{k \neq i} U_k(t) \right) / (n - 1).
\]  

(A2)

At the third layer of LCA, the relative advantage/disadvantage of each option is computed as

\[
I_i = \sum_{j \neq i} V(d_j) + I_0.
\]  

(A3)

In Equation A3, \(d_j\) is the advantage or disadvantage of Option \(i\) to Option \(j\) on the active dimension, \(V\) is a nonlinear value function with loss aversion, and \(I_0\) is a positive constant that promotes the alternatives in the choice process, namely, prevents the \(I_i\) of inferior options (i.e., with very low \(\sum_j V(d_j)\)) from being too negative.

In the fourth layer, the preference states evolve in DFT as

\[
P_i(t + 1) = v_i(t) + \sum_j s_{ij} \times P_j(t) + \xi(t),
\]  

(A4)

and for LCA according to the following equation:

\[
P_i(t + 1) = I_i(t) + \sum_j s_{ij} \times P_j(t) + \xi(t),
\]  

(A5)

with \(\xi\) standing for additive noise.

(Appendices continue)
Appendix B

Distance-Dependent Inhibition Model in Decision Field Theory

In this appendix, following decision field theory author recommendations (J. Bussemeyer, personal communication, November 4, 2009), we present a more sophisticated distance metric defined on the two new dimensions of indifference and dominance.

To illustrate how this distance metric operates, we assume two options that are characterized in two attributes, economics and quality: \( A = (E_1, Q_1) \) and \( B = (E_2, Q_2) \).

Then, we define \( (\Delta E, \Delta Q) = (E_1 - E_2, Q_1 - Q_2) \) and \( (\Delta I, \Delta D) = ([\Delta Q - \Delta E], [\Delta Q + \Delta E]) \), with \( \Delta I \) and \( \Delta D \) standing for the indifference and dominance directions, respectively.

The decision field theory (DFT) steady state solution is given as

\[
\begin{align*}
P(I) & = \left[ \begin{array}{ccc}
1 - \lambda & -\alpha & 0 \\
-\alpha & 1 - \lambda & 0 \\
0 & 0 & 1 - \lambda \\
\end{array} \right] \times \\
& \times \left[ \begin{array}{ccc}
-0.5x - 0.5y \\
x - 0.5y \\
y - 0.5x \\
\end{array} \right],
\end{align*}
\]

thus Option B is the winner. In general, for any leak parameter \( \lambda \) and if we assume that \( A(0, 0), B(\alpha, \beta), \) and \( C(\gamma, \delta) \), we can outside the range of inhibition and A and B interacting with inhibition equal to \( \alpha \), the steady state preferences for the three options are given as

\[
\begin{align*}
P(I) & = \left[ \begin{array}{ccc}
1 - \lambda & -\alpha & 0 \\
-\alpha & 1 - \lambda & 0 \\
0 & 0 & 1 - \lambda \\
\end{array} \right] \times \\
& \times \left[ \begin{array}{ccc}
-0.5x - 0.5y \\
x - 0.5y \\
y - 0.5x \\
\end{array} \right],
\end{align*}
\]

The psychological distance between Options \( i \) and \( j \) is defined as

\[
D_{ij} = \Delta I^2 + b \times \Delta D^2,
\]

with \( b > 1 \) being the weight on the dominance direction.

The Gaussian inhibition function is defined as

\[
s_{ij} = \delta_{ij} - \phi_2 \times e^{-b_i x_i^2},
\]

with \( \varphi_1 \) and \( \varphi_2 \) being the parameters of the Gaussian function and \( \delta_{ij} \) being 1 when \( i = j \) (self-feedback connection) and 0 when \( i \neq j \) (lateral inhibition).

Appendix C

Reversal of Dominance in Decision Field Theory

The decision field theory (DFT) steady state solution is \( P = (I - S)^{-1} \times V \), where \( S \) corresponds to the connectivity matrix and \( V \) to the valence vector of the choice options.

For the choice scenario in Figure 6a in the main text, \( A = (0, 0), B = (0.2, 0.2), \) and \( C = (1, 1) \), we get

\[
P = \left[ \begin{array}{ccc}
227.2 & -216.9 & 0 \\
-216.9 & 227.2 & 0 \\
0 & 0 & 20 \\
\end{array} \right] \times \\
\times \left[ \begin{array}{ccc}
-0.6 \\
-0.3 \\
0.9 \\
\end{array} \right] = \left[ \begin{array}{ccc}
-71.2 \\
62 \\
18 \\
\end{array} \right],
\]

To prevent \( P(B) > P(C) \), we need

\[
\frac{1}{2} \alpha x + \alpha y - 2x + y + 2\lambda x - 2\lambda y < \frac{1}{2} - 2y + x
\]

or

\[
\alpha > f(x, y)(\lambda - 1)
\]

Figure C1. The boundary of inhibition above which a reversal of dominance happens: B wins over C for the triple \( A(0, 0), B(x, y), \) and \( C(k, k) \), as a function of the attribute values \( x \) of the mediocre and dominated Option \( B(x, y) \) and for three different levels of \( k \) (C[1, 1], C[1.5, 1.5], and C[3, 3]). The star on the solid line (\( k = 1 \)) indicates the critical inhibition boundary for the choice scenario of Figure 6a in the main text.

(Appendices continue)
with

\[ f(x, y) = \frac{1}{2}x + y - \frac{(13x^2 - 34xy + 25y^2)}{-2y + x}. \]  

(C2)

Note that, for the choice scenario where C is outside the range of inhibition and A interacts with B, the stability is maintained for

\[ \alpha = \lambda - 1. \]  

(C3)

Therefore, the dominance reversal boundary and the instability boundary coincide only when \( f(x, y) = 1 \). In Figure C1, we can see the boundary of inhibition for dominance preservation for the triple A(0, 0), B(x, y), and C(k, k). We observe that the critical level of inhibition above which the dominance order is reversed increases as \( x \) and \( k \) are teased apart and decreases when they are brought closer.

For the scenario above (A[0, 0], B[.2, .2], C[1, 1]) and \( \lambda = .95 \), the critical inhibition boundary derived from C1 is \( \alpha = -.043 \) (see also the red star in Figure C1). The parameters that give inhibition between A and B below the critical boundary are the same as the optimized ones, except that \( \varphi_2 \) is changed from .05 to .044 (see Appendix B for an interpretation of the parameters) and noise is increased to \( \sigma = .2 \). Crucially, the leaky parameter was maintained to \( \lambda = .95 \). Under these new parameters, decision field theory still predicts the compromise and similarity effects (the attraction effect is robustly predicted for both parameter sets). The magnitude of these effects is weaker as compared to the ones given from the optimized parameter set that did not control for the reversal of dominance (see Table C1).

However, we have shown that it is possible to prevent the reversal of dominance and at the same time maintain the predictions for the three effects. Therefore, we recommend that future parameter estimations of the DFT model should predict the three reversal effects but should also be constrained by the prevention of the reversal of dominance.

Table C1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Effect</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \varphi_2 )</td>
<td>( \lambda )</td>
<td>Noise</td>
</tr>
<tr>
<td>.050</td>
<td>.95</td>
<td>.05</td>
</tr>
<tr>
<td>.044</td>
<td>.95</td>
<td>.20</td>
</tr>
</tbody>
</table>

Postscript: Contrasting Predictions for Preference Reversal

Marius Usher
Tel Aviv University and Birkbeck College

Konstantinos Tsetsos and Nick Chater
University College London

Hotaling, Busemeyer, and Li (2010) provided a valuable reply to the challenges we raised for the decision field theory (DFT) account of preference reversal in multiattribute choice. We agree with their observation that with the addition of an internal stopping rule—where a decision is reached when the first choice unit reaches a response criterion—the model is more stable and less subject to violations of dominance. Indeed, in its present form, DFT captures most existing data on preference reversals, and its limitations (due to linearity) have the virtue of facilitating analytical calculations. It is therefore interesting to contrast DFT and alternative accounts of preference reversals (e.g., leaky competing accumulators [LCA; Usher & McClelland, 2004] or the context-dependent advantage model [Tversky & Simonson, 1993]). This note builds on the improved clarity of DFT mechanisms resulting from this exchange and highlights predictions that could distinguish between competing explanations and drive further experimental research. We also note common aspects of DFT and LCA and draw implications for theories of decision making.

Consider first the attraction effect. In DFT, this is a contrast effect and is conditioned on the decoy being close to the target (within the inhibition range in the dominance direction). As the inhibition function is relatively sharp (decreasing with distance, \( x \), in the dominance direction, as \( \exp[-x^2] \)), it yields a sharp boundary for the attraction effect, in contrast to the more graded predictions obtained in LCA or other theories based on gradual value functions with loss aversion (see Figure 5 in Tsetsos, Usher, & Chater, 2010). Furthermore, these models make opposite predictions about the magnitude of the attraction effect for decoys of the type of D1, D2 (see Figure 1 in Hotaling et al., 2010), LCA and the context-dependent advantage model predict the opposite: \( P(\text{target|D1}) = .68 \) and \( P(\text{target|D2}) = .64 \). The reason for this difference is that the DFT inhibition function is higher for D2 (as this decoy is closer in the dominance direction), while, for the LCA, the magnitude of the effect depends on the relative distance of the decoy from the two competiting options (and the D2 decoy slightly dominates the target in one dimension, conferring a disadvantage). Finally, DFT differs from LCA in predicting that the attraction effect reverses when a decoy to the decoy is introduced (see Table 2 in Hotaling et al., 2010) in a four-option choice scenario. This occurs because the attraction effect in DFT is due to the decoy with negative valence boosting the activation of the target (via negated inhibition); this boost reverses with the intro-