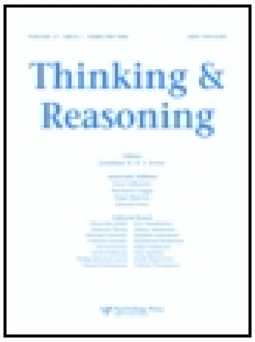


Thinking & Reasoning

Routledge
Taylor & Francis Group

ISSN: 1354-6783 (Print) 1464-0708 (Online) Journal homepage: http://www.tandfonline.com/loi/ptar20

The dynamics of bidirectional thought

Sudeep Bhatia

THINKING & REASONING, 2016
http://dx.doi.org/10.1080/13546783.2016.1187205

Routledge
Taylor & Francis Group

The dynamics of bidirectional thought

Sudeep Bhatia

Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

ABSTRACT
High-level judgement and decision-making tasks display dynamic bidirectional relationships in which salient cues determine how responses are evaluated by decision-makers, and these responses in turn determine the cues that are considered. In this paper, we propose Kosko's bidirectional associative memory (BAM) network, a minimal two-layer recurrent neural network, as a mathematically tractable toy model with which the properties of existing bidirectional models, and the behavioural implications of these properties, can be studied. We first derive results regarding the dynamics of the BAM network, and then show how these results can be used to provide an analytically sound explanation for a number of important findings, including coherence shifts in judgement and choice, anchoring effects, and reference point effects.

ARTICLE HISTORY Received 22 September 2015; Accepted 29 April 2016

KEYWORDS Judgement and decision-making; bidirectionality; constraint satisfaction; neural networks

Introduction

Bidirectionality is an important property of high-level cognition. The available evidence in a judgement task, for example, does not only determine the decision-maker's conclusions; the use of evidence is itself affected by whether it supports the conclusions that are being considered by decision-makers (Glöckner & Betsch, 2008; Glöckner, Betsch, & Schindler, 2010; Glöckner, Hilbig, & Jekel, 2014; Glöckner & Hodges, 2011; Holyoak & Simon, 1999; Kostopoulou, Mousoulis, & Delaney, 2009; Simon & Holyoak, 2002; Simon, Pham, Li, & Holyoak, 2001; Simon, Snow, & Read, 2004). This relationship is dynamic, and the use of relevant information can change over the time course of judgement, so as to cohere with emerging beliefs. Similar bidirectional relationships are also at play in preferential decision-making, with salient choice alternatives determining the attributes that are attended to, and subsequently biasing preferences and final decisions (Bhatia, 2013; Bond, Carlson, Meloy, Russo, & Tanner, 2007; Brownstein, 2003; Carlson & Pearo, 2004; Dekay,

CONTACT Sudeep Bhatia ✉ bhatiasu@sas.upenn.edu

Patiño-Echeverri, & Fischbeck, 2009; Russo, Carlson, & Meloy, 2006; Russo, Carlson, Meloy, & Yong, 2008; Russo, Medvec, & Meloy, 1996; Russo, Meloy, & Medvec, 1998; Simon, Krawczyk, Bleicher, & Holyoak, 2008; Simon, Krawczyk, & Holyoak, 2004).

As an example of this phenomena, consider the legal judgement task used by Holyoak and Simon (1999). In this task, participants were asked to choose one of two possible verdicts in a case involving six arguments. Each of these six arguments supported one verdict and opposed the other, and participants were asked to evaluate these arguments at varying stages of the judgement task. Holyoak and Simon found that prior to the decision, ratings of the validity of these arguments were not correlated, that is, participants' evaluations of one argument did not predict their evaluation of other arguments. But after the decision was made, the ratings of the arguments changed to cohere with both the emerging verdict and other arguments supporting the verdict, so that participants were more likely to rate all the arguments supporting their final verdict positively and to rate all the arguments opposing their final verdict negatively. These changes illustrate the bidirectional nature of judgement. Evidence affects decision-makers' conclusions, and these conclusions in turn influence how decision-makers use this evidence.

Theoretical approaches to studying bidirectional judgement and decision-making (e.g., Glöckner & Betsch, 2008; Glöckner et al., 2014; Guo & Holyoak, 2002; Holyoak & Simon, 1999; Spellman, Ullman, & Holyoak, 1993; Simon, Snow, et al., 2004) involve two-layered neural networks with recurrent connections. These types of networks are highly complex and, as a result, theoretically opaque. Analytical results regarding their emergent dynamics do not exist, and many of the key properties of bidirectional thought—and their implications for judgement and decision-making—are yet to be fully understood.

In this paper, we propose the bidirectional associative memory (BAM) network (Kosko, 1987, 1988) as a mathematically tractable tool for studying the properties of existing models, and for rigorously understanding the implications of bidirectional thought. BAM is a two-layered neural network, whose structure closely mimics existing models. But unlike these models, it is simple enough that its dynamics can be mathematically formalised. We begin by analysing some of these dynamics, and then apply our insights regarding these dynamics to predict human behaviour. We find that these dynamics allow our model to formally account for a range of well-known empirical results, including many results not currently attributed to bidirectional thought.

One important set of findings that we consider involve the types of coherence shifts in cue assessment discussed above, where bidirectional relationships exist between beliefs about conclusions and the evaluation of the evidence that supports these conclusions (Glöckner et al., 2010; Holyoak &

Simon, 1999; Kostopoulou et al., 2009; Simon, Krawczyk, et al., 2004; Simon et al., 2001). A related set of findings pertain to anchoring effects, in which the conclusions under consideration are ordered on a numerical scale. In these settings, asking people to consider a numerical anchor prior to deliberating can bias their evaluation of evidence and subsequently lead to final responses that are closer to the anchor than the correct response (Chapman & Johnson, 1994, 1999; Mussweiler & Strack, 1999; Tversky & Kahneman, 1974). A third set of findings that we consider involve coherence shifts in preferential choice, in which attribute ratings change over the time course of the task, so that attributes present in preferred choice alternatives are more likely to be rated as important or valuable near the end of the decision relative to that at the start (Bond et al., 2007; Carlson & Pearo, 2004; Dekay et al., 2009; Russo et al., 1996, 1998, 2006, 2008; Simon et al., 2008; Simon, Krawczyk, et al., 2004). A final application that we consider is reference dependence, in which making certain choice objects particularly salient can draw attention towards their component attributes, increasing the preference for these objects (Ashby, Dickert, & Glöckner, 2012; Carmon & Ariely, 2000; Johnson, Häubl, & Keinan, 2007; Nayakankuppam & Misra, 2005; Willemsen, Böckenholt, & Johnson, 2011). All of these settings are examples of bidirectionality as evidence influences the conclusions that people consider and these conclusions in turn influence the activation and the evaluation of evidence (in the case of judgements), and attention to attributes influences preferences for alternatives which in turn influence attribute activation (in the case of choice).

We show that, in all of these settings, the dynamical properties of the BAM network play a key role in analytically predicting and explaining behaviour. As these dynamical properties are shared with other, more complex, existing models, our results suggest that these existing models are also able to generate these behaviours. Overall, our results show that the BAM framework provides a parsimonious but mathematically structured approach to studying the properties of bidirectional models, and illustrate how these properties can be used to understand behaviour across a number of different psychological domains.

Model and key properties

Existing approaches

In this paper, the term *bidirectionality* refers to the dynamic relationship between mental objects that form the responses in various high-level cognitive tasks, and the informational cues that guide the use of these responses in these tasks. Response objects can include conclusions in judgement tasks and choice alternatives in preferential decision-making tasks, and the informational cues corresponding to these response objects can include evidence and choice attributes. In both of these domains, beliefs regarding different conclusions and preferences over different alternatives are both influenced by corresponding informational cues, and themselves influence the use of these cues. That is, salient evidence and choice attributes determine beliefs and preferences, which in turn affect the evidence and choice attributes that are attended to.

The effect of informational cues on final responses is almost always a property of rational decision-making, whereas the reverse relationship is often seen as the cause of observed deviations from rational decision-making. Indeed, the coherence shifts in judgement and choice, associated with these reverse relationships, correspond to a type of inconsistency that cannot be accommodated by rational models (Glöckner et al., 2010; Holyoak & Simon, 1999; Russo et al., 1996, 1998; Simon, Krawczyk, et al., 2004). Likewise, both anchoring effects in judgement and reference point effects in choice generate a dependence on salient response options, which is typically considered to be irrational (Tversky & Kahneman, 1974, Tversky & Kahneman, 1991).

While the most popular approaches to understanding judgement and decision-making are unidirectional, with Bayesian inference, for example, using priors and likelihood ratios to derive posterior probabilities, there has, nonetheless, been considerable theoretical work on modelling bidirectional processing and its effect on high-level cognition and behaviour. Perhaps, the most common approach to this problem involves explanatory coherence (Thagard, 1989). Theories of explanatory coherence propose that coherence—that is, mutual support through explanatory relations—plays a key role in evaluating evidence and conclusions. According to this approach, a conclusion is coherent with evidence that explains it, and incoherent with evidence that contradicts it. Judgement involves the selection of the conclusion, and accompanying explanatory evidence, that are together the most coherent.

The explanatory coherence framework has been implemented as a constraint satisfaction problem in a connectionist network, in which pairs of coherent propositions share a bidirectional excitatory connection, and pairs of incoherent propositions share a bidirectional inhibitory connection. Networks such as these have been applied, in varying forms in domains as diverse as attitude formation and change (Monroe & Read, 2008), stereotype effects (Kunda & Thagard, 1996), social reasoning (Read & Marcus-Newhall, 1993), analogical mapping (Holyoak & Thagard, 1989), dissonance reduction (Shultz & Lepper, 1996), preferential choice (Guo & Holyoak, 2002; Simon et al., 2004), and everyday judgement (Glöckner & Betsch, 2008; Glöckner et al., 2014).

As an example of this type of approach, let us briefly explore the Co3 model, a variant of Thagard's (1989) ECHO network proposed by Spellman et al. (1993), and further applied to model legal judgement by Holyoak and Simon (1999). Holyoak and Simon's (1999) application simulates human judgements in the legal setting outlined in the Introduction section. It involves network units for the two verdicts, and for each of the arguments, with symmetric positive connections between the verdicts and their supporting arguments, and symmetric negative connections between the verdicts and various conflicting arguments. Both network connections and node activation are assumed to be continuous. The Co3 model adopts the main assumption of other coherence maximising connectionist models; that is, symmetric bidirectional connectivity between units representing verdicts and arguments. As a result of this, the spread of activation in the Co3 model appears to lead to stable states that maximise local coherence between these units, and these final stable states can be biased by altering the starting activation states of the network. Both these properties allow the Co3 model to describe observed human behaviour in Holyoak and Simon's (1999) experiments (see also Simon et al., 2001; Simon, Krawczyk, et al., 2004; Simon et al., 2004; Simon et al., 2008).

Another prominent model is the parallel constraint satisfaction (PCS-DM) model (Glöckner & Betsch, 2008), which has been successfully applied to multi-cue judgement (Glöckner & Betsch, 2012; Glöckner & Hodges, 2011; Glöckner et al, 2014). This model also has two layers, with one layer corresponding to responses and the other corresponding to cues, and additionally features recurrent connections between these layers, corresponding to the degree of support a cue provides to a response. This model also features negative connections between different response options. Again both the cue–response associations and the cue activations in this model are continuous, and the model deliberates through coherence maximisation. Decisions are made based on response activation once the network stabilises.

The Co3 model and PCS-DM present an important step towards understanding the implications of bidirectional processing in judgement and decision-making. They illustrate the general behavioural patterns that can be generated by bidirectional processing, and provide a template for studying these patterns using formal models. That said, they are analytically intractable and thus theoretically opaque. For example, while many of the results these models generate depend on the effects of encoded memories (that characterise, for example, stored cue–response relationships) and exogenous inputs (that determine, for example, starting activation states) on responses, the precise nature of these relationships is not easily discernible. It is, thus, impossible to characterise the key features of bidirectional processing that are necessary to explain observed findings. Additionally, the behavioural implications of bidirectionality, in these models, are almost always illustrated using computational simulations, a methodological choice that is sensitive to a variety of highly specific modelling decisions. Subsequently, the generality of the results of these models is not always guaranteed. Incidentally, these models also feature the converse problem. As the properties of these models have not been formally characterised, it is entirely possible that they are able to explain findings that are not currently attributed to bidirectional processing; that is, it is possible that these models are more powerful than we think they are.

There is also the question of emergent dynamics. Bidirectional models involve fairly complex transitions between activated network states. These transitions characterise spreading activation in the network, and the sequence of the decision-maker's thoughts when judging different conclusions or deciding between different choice alternatives. Currently, bidirectional models do not provide any insights regarding the nature of bidirectional state transitions, and subsequently the type of spreading activation and sequence of thoughts that these transitions entail. Relatedly, despite stability being a fundamental requirement for coherence maximisation and constraint satisfaction, it is not always clear whether these state transitions guarantee the existence of final stable activation states for all types of encoded memories and all types of exogenous inputs. Does the deliberation process always terminate, or can activation continue spreading indefinitely?

It is important to note that these criticisms are unrelated to model under specification. The network structure that underpins these models is well described and these models can be applied to behaviour just like any other cognitive model. It is merely the case that unlike simpler unidirectional models (e.g., linear models or heuristics), the complexity of Co3 and PCS-DM makes it very difficult to fully specify their properties analytically. Analytical intractability of this form is unavoidable. These networks involve continuous activation functions and continuous weights, which allow for a very large number of possible relationships, and subsequently patterns of activation, for any two network units. With these issues in mind, a simpler approach to studying bidirectional judgement and decision-making may be desirable. Such an approach could retain the key structure of these existing bidirectional models, while limiting the number of different activation states and connection strengths possible in the network, thereby allowing for easier mathematical analysis of the model, and more concrete insights regarding the properties and implications of existing models of bidirectional processing, and bidirectional cognition more generally.

Bidirectional associative memory

The BAM network provides one such approach to studying bidirectionality. BAM is a minimal non-linear feedback heteroassociative memory network, introduced by Kosko (1988). It has two layers, and nodes in these two layers have symmetric bipolar connections with each other. There are no connections between two nodes in any one layer. Additionally, each node has a threshold activation function, with an activation state of either 0 or 1. Processing in the BAM network begins when nodes in one of the two layers are activated. Node updating proceeds sequentially between layers, and synchronously within layers, and processing terminates when the network stabilises. Overall, BAM generalises the autoassociative Hopfield network, which BAM resembles when both layers have the same number of nodes, and node updating within each layer is asynchronous.

We can adopt the BAM network as a toy model for a judgement or decision-making task involving N response objects based on M informational cues. Responses and cues can represent conclusions and evidence in a judgement task, or choice alternatives and attributes in a preferential decision-making task. Here, we assume that the N responses have a localist representation on a response layer, and that the M informational cues have a localist representation on a cue layer, with each of the N nodes in the response layer corresponding to each of the feasible responses and each of the M nodes in the cue layer corresponding to each of the relevant informational cues. The activation of the node corresponding to cue j in the cue layer, at time t, can be written as $c_j(t)$, and the activation of the node corresponding to response i in the response layer, at time t, can be written as $r_i(t)$.

Due to the simplicity of the BAM network, the relationship between the responses and the cues is assumed to be bipolar and symmetric, with each cue and response pair being either positively or negatively related to each other. If cue j has a positive relationship with response i, we specify a symmetric connection weight $w_{ij} = +1$ between their corresponding nodes, and if cue j has a negative relationship with response i, we specify a symmetric connection weight $w_{ij} = -1$ between their corresponding nodes. No connections exist between any two response nodes or any two cue nodes. At a given time t, the activated nodes in the response layer first send inputs, weighted by the strength of connection w_{ij}, into the cue layer. This affects the activation of the nodes in the cue layer. The activated nodes in the cue layer subsequently send inputs weighted by w_{ij} into the response layer, affecting the activation of the response nodes at $t + 1$, at which point the process repeats itself.

In addition to the inputs from the response layer, we will assume that the nodes in the cue layer receive constant exogenous inputs with strength $I^c{}_j(t)$, set to $I^c{}_j(t) = 1$ for all t and all j, unless otherwise specified. In some settings, we will also assume that some nodes in the response layer receive temporary inputs $I^r{}_i(t) = 1$ at the start of the decision process, $t = 0$. These inputs determine starting activation states in the response layer. As these inputs are temporary, we will restrict $I^r{}_i(t) = 0$ for for all i and for all $t > 0$.

In addition to this, we will assume that all of the nodes in our network have the same binary activation function, with a threshold at zero. With this assumption, we can write the activation functions of any response node *i* and any cue node *j* at time *t*, as

$$r_i(t) = \begin{cases} 1 & \text{if } \sum_{j=1}^{M} w_{ij} \cdot c_j(t-1) + I_i^r(t-1) > 0 \\ 0 & \text{if } \sum_{j=1}^{M} w_{ij} \cdot c_j(t-1) + I_i^r(t-1) \leq 0 \end{cases},$$

$$c_j(t) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} w_{ij} \cdot r_i(t) + I_j^c(t) > 0 \\ 0 & \text{if } \sum_{i=1}^{N} w_{ij} \cdot r_i(t) + I_j^c(t) \leq 0 \end{cases}.$$

As with existing models of bidirectional processing, we will assume the following: (1) The responses that are activated when the network stabilises form the decision-maker's responses in the judgement or decision-making task; (2) the cues that are activated when the network stabilises determine final assessments of cue validity and form the decision-makers' justification for their final responses; (3) the patterns of activation of the response and cue nodes while the BAM network settles capture the trajectory of decision-maker's thoughts during the time course of the task; and (4) the time taken for the network to stabilise is proportional to the time taken by decision-makers to give their final responses. The network is stable at time *t*, if there are no further changes to the network's activation states after *t*, that is, if $r_i(t) = r_i(t+1)$ and $c_j(t) = c_j(t+1)$ for all *i* and *j*.

This paper will analyse the properties of the BAM network, and use these properties to explain behaviour in judgement and decision-making tasks. Many of these properties pertain to the dependence of BAM's dynamics on the memories—that is, cue−response relationships—encoded into the BAM network, and for this, a convenient way to represent the network's memory structure will be useful. We can write the set of all *M* cues as *C*. Subsequently, the set of all cues with a positive relationship with response *i* can be written as $C_i = \{\text{cue}j|w_{ij} = +1\}$, and the set of all cues with a negative relationship with response *i* can be written as $C_i^c = C \backslash C_i = \{\text{cue } j|w_{ij} = -1\}$. The total number of cues that are positively related to response *i* can be written as $|C_i|$, and the total number of cues that are negatively related to response *i* can be written as $|C^c{}_i|$. Finally, the set of cues with a positive relationship with both response *i* and *i*′ can be written as $C_i \cap C_{i'}$, and the set of cues with a positive relationship with either response *i* or *i*′ can be written as $C_i \cup C_{i'}$. The same

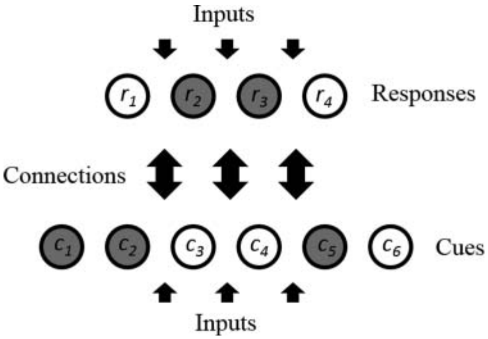


Figure 1. Overview of BAM model. The model contains two layers, corresponding to cues and responses, with bidirectional connections between the layers. Both layers can receive exogenous inputs. This example features six cues and four responses.

notation holds for the set of all N responses, R, the set of responses with a positive relationship with cue j, $R_j = \{\text{response } j | w_{ij} = +1\}$, and various compositions of these sets.

The BAM model presented here (and illustrated in Figure 1) has symmetric connections between its two layers, thereby retaining the key structure of models such as Co3 and PCS-DM described above (Glöckner & Betsch, 2008; Holyoak & Simon, 1999). Beyond this, BAM simplifies these model by only allowing two activation states for each node (on or off), and two different connection strengths between each pair of cue and response nodes (+1 or -1), and additionally avoiding connectivity within its two layers. Its dynamics are tractable, and analytical results regarding these dynamics can be used to characterise the key properties of Co3 and PCS-DM and other existing bidirectional models, as well as to explain behaviour in a variety of judgement and decision tasks. An outline of the properties of Co3, PCS-DM, and BAM is presented in Table 1.

Summary of model properties

This section outlines the key properties of the BAM network (additional details are provided in the Appendices). Later on, we will use these properties to predict behavioural effects such as coherence shifts in judgement and choice, anchoring effects, and reference point effects.

Table 1. Comparison of the properties of the BAM network and related models.

Property	Co3	PCS-DM	BAM
Recurrent bidirectional connections	X	X	X
Continuous activation states	X	X	–
Continuous weights	X	X	–
Within-layer connections	X	X	–

Now, one of the most important properties of bidirectional networks (and related neural and associative networks) pertains to the spread of activation from one response node to another. Understanding the spread of activation across responses, and the state transition rules that characterise this spreading activation, is necessary for fully characterising the BAM's behaviour. As shown in the Appendices, the spread of activation in the BAM network depends primarily on cue overlap, and activating one response activates a second response only if they overlap on enough cues. Thus, in a two-response setting, if response 1 is the only activated response, then response 2 activates at the next time period if and only if $|C_1 \cap C_2| > |C_1 \cap C^c{}_2|$. Intuitively, if the cues that are positively related to an activated response also support a second response, then the second response will activate.

Cue overlap also determines response deactivation. If the cues that are activated with a certain response are, on average, negatively related to a second response, then that second response will turn off. Thus, if the activation of response 1 leads to the activation of response 2, then $C_1 \cup C_2$ is the set of cues that are activated. Subsequently, response 1 will turn off at the next time period if the cues that are negatively related to response 1 make up the majority of this set, that is, if $|C^c{}_1 \cap [C_1 \cup C_2]| \geq |C_1 \cap [C_1 \cup C_2]|$, which simplifies to $|C^c{}_1 \cap C_2| \geq |C_1|$.

Similarity assessments are also often based on cue (or feature) overlap, and this implies that similarity may be closely related to bidirectional processing. Consider, for example, the well-known asymmetry of feature-based similarity assessment. More people agree with the statement "North Korea is similar to China" than with the statement "China is similar to North Korea". Phenomena such as this are typically explained by the dependence of feature attention on the focal category, as in Tversky's (1977) contrast model. The BAM network presents a way to formalise this dependence, and is able to generate this asymmetry. Assume that the focal category is activated at the start of the similarity assessment, and that two categories are judged as similar if activation spreads from the focal category to the non-focal category, and not similar otherwise. In a two-response setting, with response 1 corresponding to the focal category and response 2 corresponding to the non-focal category, this happens only if $|C_1 \cap C_2| > |C_1 \cap C^c{}_2|$. For assessments of similarity when response 2 is the focal category, we require an alternate set of conditions, that is, we need $|C_2 \cap C_1| > |C_2 \cap C^c{}_1|$. It is possible for this second condition to be violated even if the first condition is not, and the settings in which this happens are precisely those in which Tversky's model yields asymmetries. Indeed, it is possible to show that the similarity assessments generated by BAM are identical to those generated by a parsimonious version of Tversky's model, and BAM can be seen as representing a special case of Tversky's model.

Could spreading activation in the BAM network continue indefinitely? The answer to this is no. Based on the analysis in Kosko (1988), we can prove that any BAM network, with any number of nodes, with any encoded memories, starting at any point, with any (constant) exogenous inputs, will necessarily stabilise, and do so in a finite number of time steps. The network is guaranteed to stabilise and the decision-maker is always able to provide a response (such claims cannot be made for more complex models). Moreover, it is also the case that a response can only be activated in a stable state if it is supported by at least half of the cues that are activated in that state, and a set of cues can only be activated in a stable state if the responses that they support are activated in that state. Intuitively, decision-makers will only select responses that are coherent with activated cues.

Finally, to fully characterise BAM's behaviour, we need to understand the effect of exogenous inputs into the network on the spread of activation in the network. We assume that temporary exogenous inputs $I^r{}_i(t) = 1$ to the response layer at the start of the decision process ($t = 0$) determine the starting activation states for the response nodes. The simplest case involves the setting in which the decision-maker has no reason to favour any one response and we have $I^r{}_i(0) = 0$. This response activation state sends no feedback into the cue layer, and the cues in the set of relevant cues C all activate at the first time period. In this setting, the responses that are supported by the majority of cues in C are the only ones that receive net positive inputs in the subsequent time period, and are the only ones that are activated. Many decision tasks do involve a response bias, that is a focal response that is activated at the start of the decision. In these settings, $I^r{}_i(0) = 1$ activates the focal response and activation spreads from this focal response to other responses, based on cue overlap and similarity. Changing the focal response affects the trajectory of spreading activation, and subsequently the final stable response.

Recall, we also assume another set of exogenous inputs into the network. These inputs, $I^c{}_j(t) = 1$, affect the cue nodes, and are persistent over the time course of the decision. These inputs ensure that cue nodes are activated even when none of the response nodes are active, and that the judgement or decision process can begin in the absence of a response bias, and continue even if all responses extinguish. They also allow us to model the effect of task-related determinants of cue salience.

Applications

Overview of applications

In the remainder of the paper, we will use the insights presented above to predict human behaviour. Particularly, we will examine two psychological representations of the network. In the first representation, nodes in the response layer will correspond to conclusions, nodes in the cue layer will correspond to evidence, and we will use the BAM network to study the properties of bidirectional processing in everyday judgement. The applications considered here will be coherence shifts in the evaluation of evidence, in which the ways in which decision-makers evaluate evidence changes with emerging beliefs (Glöckner et al., 2010; Holyoak & Simon, 1999; Kostopoulou et al., 2009; Simon, Krawczyk, et al., 2004; Simon et al., 2001), and anchoring effects, in which making a certain numerical conclusion more salient influences the way in which evidence is evaluated and leads to final responses closer to this conclusion than optimal (Chapman & Johnson, 1994, 1999; Mussweiler & Strack, 1999; Tversky & Kahneman, 1974). Both of these phenomena are examples of bidirectionality as evidence influences the conclusions that people consider and these conclusions in turn influence the activation and the evaluation of evidence. In the second representation, nodes in the response layer will correspond to choice alternatives, nodes in the cue layer will correspond to choice attributes, and we will use the BAM network to study the properties of bidirectional processing in preferential decision-making. The applications considered here will be coherence shifts in the evaluation of attributes, in which the evaluation of attributes changes with emerging preferences (Bond et al., 2007; Carlson & Pearo, 2004; Dekay et al., 2009; Russo et al., 1996, 1998, 2006, 2008; Simon, Krawczyk, et al., 2004; Simon et al., 2008), and reference dependence effects, in which making certain choice options reference points increases the activation of their component attributes, and ultimately leads to a preference for these options (Ashby et al., 2012; Carmon & Ariely, 2000; Johnson et al., 2007; Nayakankuppam & Misra, 2005; Willemsen et al., 2011).

Again, the insights presented in these sections should also hold for more complex models such as Co3 and PCS-DM (Glöckner & Betsch, 2008; Holyoak & Simon, 1999). Indeed, some of these models have already been shown to successfully predict coherence shifts in judgement and choice. Likewise, these models have been suggested as possible explanations for anchoring (Glöckner & Englich, 2015; Russo, 2010), though these explanations have not yet been tested. What is unique about using BAM to study these effects, however, is its analytical tractability. This allows us to fully specify the features of bidirectionality that underpin these effects, and to clearly characterise the settings in which these effects are likely to emerge. This is necessary for a complete and rigorous understanding of the implications of bidirectionality for findings like coherence shifts, anchoring, and reference dependence.

Coherence shifts

Let us first use our model to capture and understand results related to coherence shifts regarding conclusions and their accompanying evidence, in judgement. As outlined in the Introduction section, Holyoak and Simon (1999) presented participants with a legal case with two possible verdicts and a series of arguments supporting the two verdicts. As expected, prior to the decision, ratings of the validity of these arguments were not correlated. But after the decision was made, and the participants reached a verdict, ratings of the arguments cohered with both the emerging verdict and other arguments supporting the verdict. More specifically, participants were more likely to rate all the arguments supporting their final verdict positively and to rate all the arguments opposing their final verdict negatively. A similar work by Simon et al. (2004) shows that prior participant preferences for a verdict, or pre-assigning participants to support a verdict, increases their chance of choosing that verdict, and also leads to accompanying coherence shifts in judgements of argument validity. In addition to this, Simon et al. (2001) showed that the above effects emerge regardless of how the legal case is presented and learnt. Relatedly, Kostopoulou, Mousoulis, and Delaney (2009) find coherence effects in medical judgement, and Glöckner et al. (2010) find these effects in judgement tasks involving ratings of cue validity.

The dynamics of BAM provide a rigorous explanation for these findings. Let us assume that the verdicts in Holyoak and Simon (1999) correspond to responses (conclusions), and that the arguments correspond to cues (evidence). We will consider a setting with two responses, responses 1 and 2 (corresponding to the two verdicts in Holyoak and Simon's experiment) and M cues, with each cue supporting one response and opposing the other (another feature of Holyoak and Simon's experiment). This means that the sets of cues supporting each of the two responses do not overlap, and also together compose the entire set of relevant cues, that is, $C_1 \cap C_2 = \emptyset$ and $C_1 \cup C_2 = C$. We will assume, for simplicity, that response 1 is correct, but that some cues, nonetheless, support response 2, that is $|C_1| > |C_2| > 0$.

Consider a setting without a salient response verdict to bias the decision, that is, with all response nodes deactivated at $t = 0$. Using the insights presented in the above sections and in the Appendices, we can prove that in this setting, all of the relevant cues, C, and none of the two responses will be activated at the start of the decision. Subsequently, when the network stabilises, response 1—the correct response—will be the only activated response, and cues that supporting this response, C_1, will be the only activated cues. Cues that do not support response 1 (i.e., those in C_2) will be deactivated. The activation of only the cues that support the final response captures the main coherence results of Holyoak and Simon (1999), with the additional caveat that the model, despite displaying coherence-based shifts in its activation of cue nodes, nonetheless, selects the response which has the majority of cues supporting it, a response that in most settings is considered correct. This type of response dynamic is shown in Figure 2.

What happens when the salience of one of the responses is manipulated experimentally? As in Simon, Krawczyk, et al. (2004), we will assume that response nodes that are exceptionally salient receive exogenous inputs and are activated at the start of the decision process. Using the insights presented

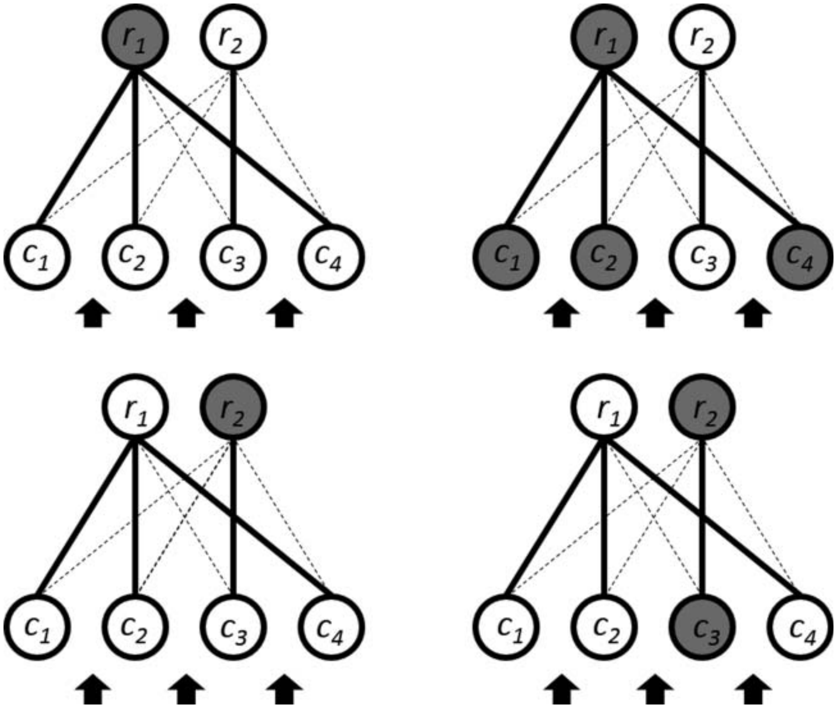


Figure 2. Network dynamics with responses biases. The top left panel displays the setting in which response 1 is activated, and the top right panel displays the resulting network stable state from this activation. The bottom left displays the setting in which response 2 is activated, and the bottom right panel displays the resulting network stable state from this activation.

in the above sections, we can prove that these settings lead to the stable activation of both the biased response node and its supporting cues, as long as there is at least one cue that supports the biased response. Thus, if response 1 is the focal response, it will characterise the stable state, along with the cues in C_1. The opposite is the case if response 2 is the focal response. This allows us to capture the results of Simon et al. (2004), in which participants preassigned to consider a verdict are more likely to choose that verdict, and believe its supporting arguments, at the end of the decision process. The network dynamics corresponding to these two settings are shown in Figure 2.

As these results are often presented as *coherence shifts* in judgement, it may be useful to briefly reinterpret the above explanations in terms of the coherence maximisation properties of the BAM network. The BAM network presented here, with two responses and a memory structure with $C_1 \cap C_2 = \emptyset$, $C_1 \cup C_2 = C$, and $|C_1| > |C_2| > 0$ has only two possible stable states, that is, two states of maximal local coherence. The first involves only the activation of response 1 and of cues in C_1, whereas the second involves only the activation of response 2 and of cues in C_2. As shown in the Appendices, for the simple setting described here, without cue overlap, the coherence level of the first stable state is $\text{Coh}_1 = |C_1|$, and that the coherence level of the second stable state is $\text{Coh}_2 = |C_2|$. In the absence of a response bias, the coherence level of the starting state is $\text{Coh}_0 = 0$, and in this setting, the network will slip into the first stable state, which has a higher coherence level than the second stable state, and is the state with maximal global coherence. If, however, there is a response bias, as with preassigned responses, the initial state of the network will slip into the closest state of maximal local coherence, which will then be the final stable state. That is, if the network begins with response 2 activated, it will stabilise in the second stable state, with response 2 activated, even though Coh_2 is less than Coh_1. Why does this happen? In order for the network to move from the activation of response 2 and cues in C_2, to the activation of response 1 and cues in C_1, the network needs to transition through either the state in which both responses 1 and 2 and the cues in C_2 are simultaneously activated, or the state in which responses 1 and 2 are deactivated and only the cues in C_2 are activated. Both of these states have a coherence of 0, thereby barring the network from transitioning through them.

As mentioned above, these results can also be understood in terms of similarity. Holyoak and Simon (1999) and Simon, Krawczyk, et al. (2004) assume that the two responses do not overlap on the cues, that is, that these responses are completely dissimilar. In these settings, activation does not spread from one response to another, ensuring that only one response remains activated once the network stabilises. The cues that are activated in the stable state are the ones that support this response (and oppose the other).

Finally, note that the results presented here are analytically sound, and hold for all possible parameter values and stored cue–response relationships with the above structure. Additionally, as can be seen in the Appendices, these results are a product of the fundamental mathematical properties of bidirectionality. Although other more complex bidirectional models have already been proposed as explanations for coherence shifts, these explanations are based on simulations and cannot be understood with the same degree of analytical depth as can be done with the BAM network.

Anchoring

Two theories of anchoring

Anchors have a powerful effect on numerical judgement. Responses in these types of tasks are systematically affected by uninformative numbers, known as anchors, which are displayed to the decision-maker prior to the judgement task. High anchors generate high responses, low anchors generate low responses, and final judgements can be manipulated by selecting the appropriate anchor.

The anchoring effect was first demonstrated by Tversky and Kahneman (1974). Tversky and Kahneman generated a random number (an anchor) between 1 and 100, and asked participants whether the percentage of African countries in the United Nations was higher or lower than this number. After this, participants were required to provide an overall estimate of the percentage of African countries in the United Nations. Tversky and Kahneman found that participants with higher random numbers in the first task listed higher percentages in the second task, despite the fact that the anchored number was randomly generated and thus completely uninformative.

While the anchoring effect is remarkably robust, and has been shown to emerge in a number of domains, the cognitive mechanisms responsible for the effect are still unclear. Tversky and Kahneman (1974) proposed that anchoring is caused by an imperfect sequential adjustment process. At each step in this process, decision-makers evaluate the validity of a particular number as a response in the judgement task. The judgement process terminates if the number under consideration appears adequate; otherwise, it moves on to the next feasible value. Anchors determine the starting point in this process, and adjustment is insufficient. Subsequently, final responses are closer to the anchor than optimal.

This explanation for the anchoring effect has been popular for many decades, and most formal models of the anchoring effect have assumed that anchoring operates through sequential adjustment (e.g., Hogarth & Einhorn, 1992; Johnson & Busemeyer, 2005). A more recent approach, however, claims that anchoring is the product of biased activation (Chapman and Johnson, 1994, 1999; Mussweiler & Strack, 1999; Mussweiler, Strack, & Pfeiffer, 2000; Strack & Mussweiler, 1997). Anchors, according to this view, increase the activation of evidence supporting the anchor. This evidence generates final answers that are closer to the anchor than optimal. Unlike the sequential adjustment theory, which assumes that decision-makers search through feasible responses in a rule-based manner, biased activation theory claims that anchors influence responses by affecting the memory processes underlying judgement.

Is anchoring caused by sequential adjustment or biased activation? Both theories are supported by a large number of empirical findings (discussed later), but neither is able to predict all of these findings by itself. In this paper, we provide a simple answer to this question. We show that these processes are not necessarily distinct: sequential adjustment emerges from the dynamics of biased activation, when biased activation is formalised using the BAM network. Anchoring, thus, is caused by both of these mechanisms simultaneously, and a large range of findings regarding anchoring and its moderators can be explained through the operations of bidirectional processing.

The emergence of sequential adjustment

We assume different numbers considered by the decision-maker as conclusions in the anchoring task are represented on the response layer, and that evidence for these numbers is represented on the cue layer. This means that the N response nodes are ordered in a sequence corresponding to the sequence of available responses. For example, when considering the percentage of African countries in the United Nations, with responses in intervals of 1%, we have $N = 100$ different responses, with response 1 corresponding to 1%, response 2 corresponding to 2%, and so on. We also assume that the anchor determines the starting state of the network, that is, at $t = 0$, the response node corresponding to the anchored number receives exogenous inputs and is activated, and all other nodes are deactivated.

We hope to show that this settling process of the BAM network in the presence of anchors resembles sequential adjustment. Before we can do this, however, we need to understand what sequential adjustment really is. Sequential adjustment is generally defined as the successive movement through the range of responses available to the decision-maker. In the simplest case, this definition imposes a form of serial processing, according to which only one response is considered at any given time. For example, when judging the proportion of African countries in the UN, decision-makers may first consider 1%. After rejecting this response, they would consider 2%. If this too is inadequate, they would move on to 3%, and so on. We consider the more general (and more realistic) case in which multiple responses can be considered at the same time. This allows decision-makers to focus on all the responses within a particular interval, such as 1%−10%, simultaneously, before moving on to the next interval in the sequence.

Such a dynamic is compatible with the general idea underlying sequential adjustment, as long as the responses activated are contiguous. Sequential adjustment does not seem to permit the simultaneous consideration of different, non-neighbouring responses. For example, decision-makers who consider both 1% and 99% simultaneously, without considering the responses between these two numbers, would not appear to be displaying sequential adjustment. We can, thus, state the first requirement for sequential adjustment. This requirement, titled *contiguous activation*, states that sequential adjustment must not involve the simultaneous activation of multiple non-neighbouring responses. Responses must be considered individually or in contiguous intervals. More formally, when the anchoring task begins with a single anchor, response activation should be such that for all t, if we have $r_i(t) = 1$ and $r_{i''}(t) = 1$, then we also have $r_{i'}(t) = 1$ for $i < i' < i''$.

Settling dynamics that display contiguous activation do not necessarily resemble sequential adjustment. It is possible for the decision-maker to consider responses in contiguous intervals at any given time, but transition across different intervals in a non-sequential manner. For example, when evaluating the proportion of African countries in the UN, decision-makers could begin by considering the interval 1%−10%, and then move to the interval 20%−30%, without considering the interval 10%−20%. We thus need an additional requirement for our definition of sequential adjustment, in order to rule out these types of dynamics. This requirement, titled *sequential transition*, states that sequential adjustment must not involve changes in activation that skip over a set of responses. Changes to response activation must be successive. More formally, for two intervals $[a, b]$ and $[c, d]$ where $a < b < c < d$, if we have $r_i(t) = 1$ for $a \leq i \leq b$ and $r_i(t) = 0$ for all other i, and $r_i(t + 1) = 1$ for $c \leq i \leq d$, then we must also have $r_i(t + 1) = 1$ for $b < i < c$.

Do the dynamics of the anchored BAM network satisfy contiguous activation and sequential transition? Not necessarily. However, with a simple assumption about the underlying memory structure, these requirements can indeed be satisfied. This assumption relates to the distribution of cue overlap across the responses. In numerical judgements, cues can seldom support two disparate responses without supporting intermediate responses. For example, when judging the proportion of African countries in the UN, any cue that supports the 10% response, and the 12% response, should, in general, support the intermediate 11% response. This property, titled *connectedness*, requires that a cue that supports two non-neighbouring responses must also support any intermediate responses, or, more formally, that for any cue j, if cue $j \in C_i$ and cue $j \in C_{i''}$, then cue $j \in C_{i'}$ for $i < i' < i''$. Memory structures displaying this property involve cues with a single, connected, interval of supported responses, whereas those that do not display this property have cues with multiple, fragmented, intervals of supported responses. While connectedness may not be satisfied in all judgement tasks, it is certainly a reasonable assumption when responses are ordered, as with the numerical scales used in anchoring experiments. Cues in these settings generally provide support for large responses, or small responses, or medium responses, or some other connected interval of responses. Very few cues provide support for a set of non-neighbouring responses, distributed sporadically across the response scale. Indeed, it is quite difficult to think of memory structures with diagnostic cues for ordered responses that do not satisfy the connectedness property.

When memory structures satisfy connectedness, then the resulting BAM network, with the anchored response activated at the start of the decision process, satisfies both contiguous activation and sequential transition. Of course, satisfying these properties does not imply that the decision-maker necessarily adjusts away from the anchor. It may be the case that the anchor is stable. If there is adjustment, however, the adjustment is guaranteed to be sequential. Anchors trigger a cascade of spreading activation in the response layer: neighbouring responses activate and deactivate consecutively. There are no jumps in response activation, nor do multiple non-neighbouring responses activate, without the activation of the intermediate responses.

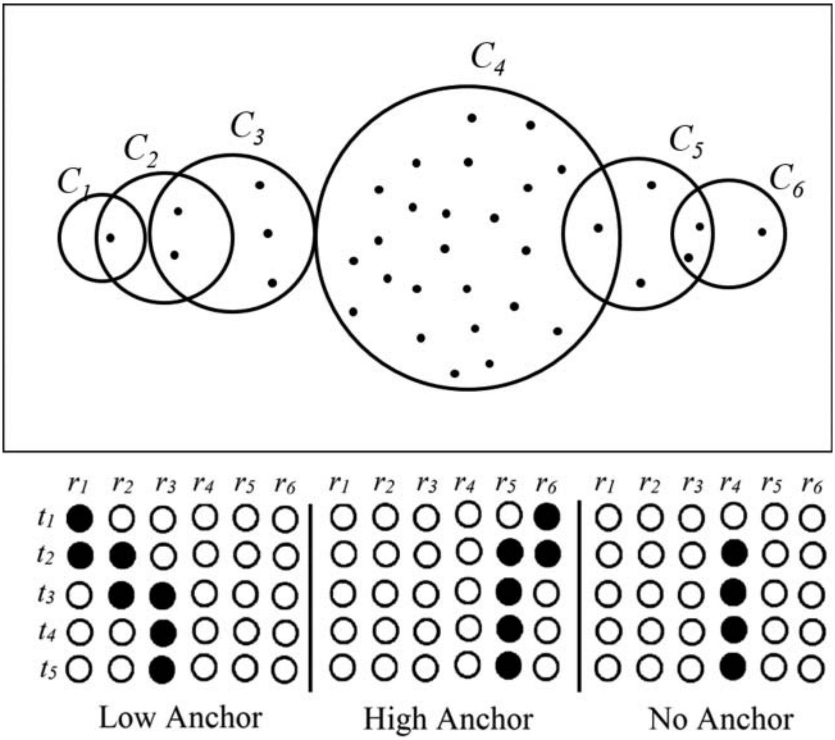
How does the connectedness property satisfy contiguous activation and sequential transitions? While the proof of this claim is in the Appendices, the intuition for it is as follows: connectedness ensures that cue overlap must be ordered, which then ensures that the spread of activation across responses must be ordered. More specifically, cues that support both the anchor and a non-neighbouring non-anchored response must also support any intermediate responses, lying between the anchor and the non-neighbouring response. Thus, if the activation of the anchor activates cues that subsequently activate non-neighbouring responses, these cues must also activate all of these intermediate responses. Subsequently, response activation at $t = 2$ must be contiguous, and any transitions that may have happened at $t = 1$ must be sequential. This intuition, however, also applies for the contiguous interval of responses activated at $t = 2$, implying that any further changes to activation after $t = 2$ must be sequential. Additionally, once a contiguous interval of responses is activated, connectedness implies that this interval cannot splinter into smaller, non-contiguous activated intervals, implying that contiguous activation must also be satisfied after $t = 2$. Mathematical induction shows that these properties then hold at all times.

Connected BAM memory structures guarantee sequential activation. But can they generate insufficient adjustment? Let us consider the case with one correct response. When the memory structure is such that two nodes lying between the anchor and the correct response do not overlap on an appropriate number of cues, the sequential adjustment process described above will be insufficient: it will not stabilise with the activation of the correct response.

The intuition for this is fairly straightforward. If, for a low anchor, there exist two response nodes, i and $i + 1$, between the anchor and the correct response, whose cue support does not overlap sufficiently, then the activation of i will not lead to the activation of cues that activate $i + 1$. As activation must be contiguous and transitions must be sequential, no nodes greater than $i + 1$ can be activated, the network will stabilise with the activation of incorrectly low responses, and the correct response will remain turned off. The same intuition holds for tasks involving a high anchor, in which the network will stabilise with the activation of incorrectly high responses, and the correct response will remain turned off.

What happens in the absence of an anchor? As discussed in the previous sections, this setting leads to the stable activation of a correct response and its accompanying cues, if such a response exists.

Figure 3 provides a demonstration of the anchoring effect as observed with the BAM network. It shows a hypothetical distribution of cue support for a sequence of responses, and the settling dynamics of the corresponding BAM network with a high anchor, low anchor, and without any anchor.



	Low Anchor						High Anchor						No Anchor					
	r_1	r_2	r_3	r_4	r_5	r_6	r_1	r_2	r_3	r_4	r_5	r_6	r_1	r_2	r_3	r_4	r_5	r_6
t_1	●	○	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○
t_2	●	●	○	○	○	○	○	○	○	○	●	●	○	○	○	●	○	○
t_3	○	●	●	○	○	○	○	○	○	○	●	○	○	○	○	●	○	○
t_4	○	○	●	○	○	○	○	○	○	○	●	○	○	○	○	●	○	○
t_5	○	○	●	○	○	○	○	○	○	○	●	○	○	○	○	●	○	○

Figure 3. Distribution of cue support, and resulting network dynamics for low, high and no anchors. In the top panel, C_i indicates the set of cues supporting response r_i. Here, even though r_4 is the correct response, having a low anchor, r_1 will lead to the network stabilising at r_3, whereas having a high anchor, r_6 will lead to the network stabilising at r_5.

The correct response in this network is response 4, and this is the stable response in the absence of an anchor. When anchored at response 6 (a high anchor), however, the network stabilises at response 5. Similarly, when anchored at response 1 (a low anchor), the network stabilises at response 3. These behaviours indicate the presence of the anchoring effect. Additionally, the settling dynamics with these anchors display sequential adjustment: response nodes activate and deactivate consecutively until the network stabilises.

These behaviours emerge because the network satisfies connectedness, which implies that the activation dynamics generated by the anchor display contiguous activation and sequential transitions, leading to sequential adjustment. Specifically, responses 1, 2 and 3 overlap on the component cues in such a way that the cues that support the first response also, on average, support the second response, and the cues that support the second response also, on average, support the third response. This means that activating response 1 leads to the activation of response 2, which then activates response 3. However, the set of cues supporting response 3 and response 4 do not overlap in this way, implying that the cascade of spreading activation begun by anchoring the network at response 1 ends with the stable activation of response 3. This generates an incorrectly low response for a low anchor. A similar property holds for response 5 and response 6, which generates an incorrectly high response for a high anchor. In the no anchor condition, response 4, which is supported by the majority of the cues in the network, and is thus the correct response, is the only one which is activated when the network stabilises.

These dynamics also emerge with larger, randomly generated memory structures. Consider a setting with $N = 100$ responses and $M = 1000$ cues. Let us randomly generate support or opposition between these cues and these responses. For each cue, we can pick a number from the normal distribution with mean 50 and variance 25, and round it to its nearest integer. We can subsequently take an interval of length 20 around this integer, to generate the set of responses supported by the cue. All other responses are opposed by the cue. Taking an interval of responses around the randomly chosen number generates a blurring in the underlying memory structure: it is seldom the case that individual cues support point estimates; rather their support is distributed across an interval of responses.

As the randomly generated memory structure satisfies connectedness, it should be able to generate sequential adjustment. Figure 4 displays the dynamics of the BAM network instantiating this randomly generated memory structure, with a high anchor, r_{100}, and a low anchor, r_1. Note that the stable responses for the two anchors are different, with the stable responses for the low anchor lower than the stable responses for the high anchor. Additionally, activation at all points of time is contiguous, and all transitions are sequential: we can observe a cascade of activation in the response layer over time, with intervals of responses activating and deactivating consecutively before finally stabilising.

Note that the dynamics observed in Figure 5 also emerge with alternate memory structures. In general, however, increasing the ratio of total responses to total cues and increasing the blurring in the cue support for the responses generate a higher likelihood of adjustment, as well as longer sequences of adjustment. This subsequently leads to weaker anchoring effects. Overall, the anchoring bias is strongest when there are many relevant cues, and each cue supports a few neighbouring responses.

Before proceeding, let us briefly reinterpret the dynamics of the anchoring effect in terms of similarity and coherence. After the anchored response is activated, the network changes activation so as to move to states with progressively higher coherence, by activating responses that are similar in terms of cue overlap to the activated network state. In networks that satisfy connectedness, neighbouring responses are necessarily more similar to each

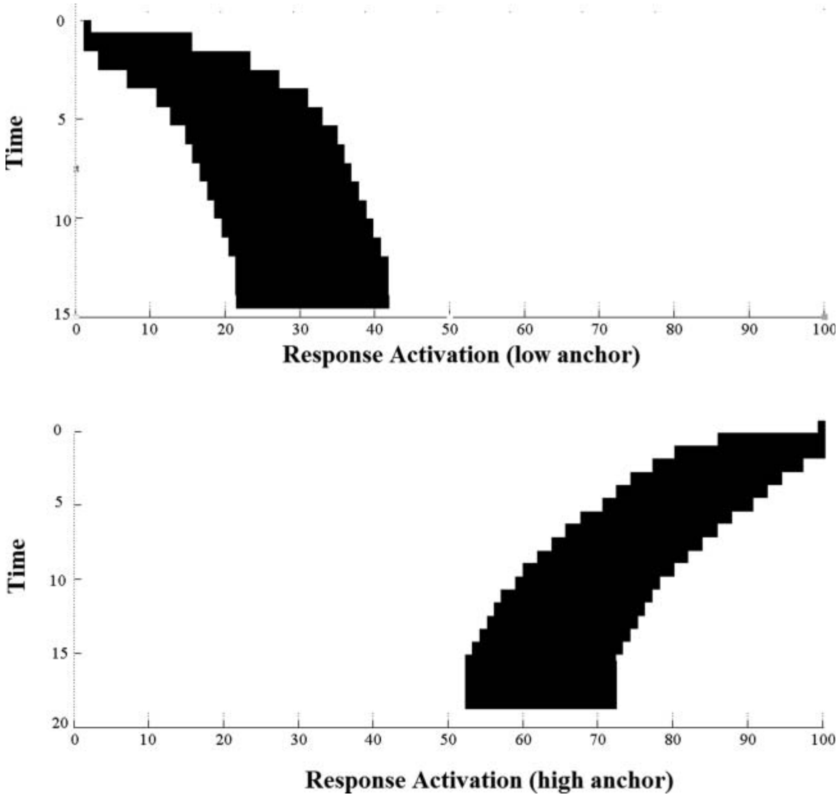


Figure 4. Network dynamics for high and low anchors, with randomly generated memory. Here, in the low anchor case, the anchored response is r_1, whereas in the high anchor case, the achor is r_{100}.

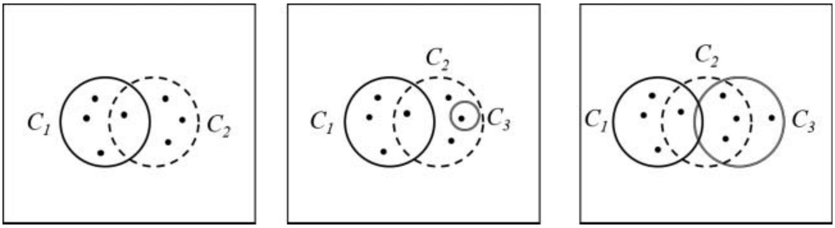


Figure 5. Examples of cue overlap that characterise reference-dependent biases in choice. The left panel corresponds to the endowment effect, the middle panel corresponds to the improvements vs. trade-offs effect, and the right panel corresponds to the advantages vs. disadvantages effect. In the middle and right panels, response 3 is assumed to be the reference point.

other than distant responses. Indeed, connectedness can be reinterpreted as a similarity-ordering property. For networks that satisfy connectedness, settling dynamics generate a cascade of coherence-increasing spreading activation, through sequences of similar responses, which resembles sequential adjustment. That is, the anchored response activates a set of similar, neighbouring numerical responses, whose coherence is higher than that of the anchored state. This, then, activates another group of similar neighbouring responses whose coherence is higher than the first group, and this continues. In many settings, this progression of increasing coherence terminates at a local coherence maximum prior to activating the correct response. This happens when two neighbouring nodes are not similar enough. In these cases, adjustment is both sequential and insufficient.

Explaining anchoring effects

Anchoring is a well-studied phenomenon and there are a large range of behavioural findings that the sequential adjustment and biased activation theories of anchoring have attempted to address. The above sections have shown that these theories are almost identical: the process assumed by one emerges directly from the process assumed by the other. This section shows how this result can explain many of the findings documented in anchoring research.

Using a lexical decision task, Mussweiler and Strack (2000) found that decision-makers identified cold-related words quicker and more accurately after temperature judgements with low anchors, and identified hot-related words quicker and more accurately after temperature judgements with high anchors. This is taken as strong evidence for the biased activation theory of anchoring, and the sequential adjustment theory is unable to account for this finding. However, the BAM framework allows for both sequential adjustment and anchor-dependent cue activation biases to emerge simultaneously: once the network settles, the cues that support the stable responses are themselves stable. If the task begins with a low anchor, then the final stable response is itself relatively low, and subsequently the cues that are stable are the ones that support this low response. The opposite is the case if the task begins with a high anchor. More formally if response a_l is the low anchor, response i_l is the final response with the low anchor, response a_h is the high anchor, and response i_h is the final stable response with the high anchor, then we would have $a_l \leq i_l \leq i_h \leq a_h$. When the network stabilises with response i_l or response i_h activated, then the cues in C_{i_l} or C_{i_h} would themselves be activated. Due to connectedness, the cues in C_{i_l} cannot be in C_{a_h} without also being in C_{i_h}, and cues in C_{i_h} could not be in C_{a_l} without also being in C_{i_l}. Thus, the cues that are activated in the low anchor condition would support lower final responses than the cues activated in the high anchor condition.

The biased activation theory of anchoring also predicts that exogenous factors that increase the activation of non-anchored cues can mitigate the anchoring effect. This has been verified by Chapman and Johnson (1999) and Mussweiler et al. (2000). Unlike sequential adjustment theory, the BAM model can explain these findings. If we assume that exogenous influences on cue attention affect the inputs, $I^c{}_j(t)$, into the cue layer at time $t = 0$, then directing attention towards cues that oppose the anchored response a, that is cues in $C^c{}_a$, leads to stronger inputs $I^c{}_j(1) > 1$ into these cues. Due to these inputs, these cues are not inhibited by feedback from the activated anchor in the response layer, as would be the case in the absence of higher inputs. Subsequently, all cues in $C_a \cup C^c{}_a = C$ are activated, and the pattern of activation on the cue layer from $t = 2$ onwards resembles the pattern observed in the absence of an anchor, causing the network to stabilise without an anchoring bias.

According to the traditional sequential adjustment theory, the background knowledge of the decision-maker should not influence the strength of the anchoring effect. Mussweiler and Strack (2000), however, find that the extent of the decision-maker's knowledge matters, with stronger anchoring effects for decision-makers with minimal knowledge of the domain in consideration. BAM can account for this effect if it is assumed that decision-makers with minimal knowledge have fewer stored cue−response associations. As spreading activation, and thus anchoring, in the BAM network depends on cue-overlap, this setting would display less cue overlap and thus less anchoring and less adjustment. Relatedly, Mussweiler and Strack (2001) find that the anchoring effect weakens if the anchor is not applied to the target, that is, if the anchor acts as a purely numeric prime rather than a semantic prime. The BAM network can easily explain this, if we assume that numeric primes do not activate BAM nodes corresponding to the target responses, as these nodes are not used to represent numbers generally (only numbers pertaining to the conclusions in question, such as the number of African countries in the United Nations).

Yet, another property of the model pertains to implausible anchors. Research by Chapman and Johnson (1994) finds that implausible anchors (anchors that are not supported by any cues) have a much weaker effect than plausible anchors. Although further work (Mussweiler & Strack, 2000) has not found this to hold consistently (see the discussion of exemplar and category knowledge below), BAM does generate this pattern of behaviour. Particularly, when implausible anchors are activated at the start of the decision process, all cues in C are suppressed (as these anchors are not supported by any cues). Subsequently, none of the response nodes activate in the next time period. This leads the network to a state identical to the starting state of the network in the absence of an anchor. Implausible anchors thus, according to the BAM model, do not generate an anchoring effect.

A fifth finding supporting the biased activation theory of anchoring pertains to the effect of multiple anchors. Sequential adjustment theory predicts that the decision-maker adjusts sequentially away from the one anchor presented in the decision task. This theory cannot make predictions for settings with multiple anchors. Switzer and Sniezek (1991) and Whyte and Sebenius (1997), however, demonstrate that multiple anchors affect judgement differently relative to single anchors. Single anchors paired with more extreme anchors generate a stronger anchoring effect than the single anchors alone, whereas single anchors paired with less extreme anchors generate a weaker anchoring effect than the single anchors alone.

BAM can account for the effect of multiple anchors. When a low anchor a_l is paired with a more extreme low anchor $a_L < a_l$, then, due to connectedness, the set of cues activated, $C_{a_l} \cup C_{a_L}$, support more low responses, relative to when a_l is activated by itself. This is because cues in C_{a_l} cannot support any responses lower than a_L without also supporting a_L and thus being in C_{a_L}. In this setting, the network will stabilise with the activation of low responses, closer to the extreme anchor, generating a stronger anchoring effect. The opposite happens when a low anchor is paired with a moderate anchor. Here, fewer activated cues support extreme responses. This leads to the ultimate stable activation of responses close to the moderate anchor, generating a weaker anchoring effect.

The results discussed above present strong evidence for the biased activation theory of anchoring. The standard biased activation theory cannot, however, provide a comprehensive account of all the moderators of the anchoring effect. Research by Reitsma-van Rooijen and Daamen (2006), for example, finds that time pressure increases the anchoring effect. This has traditionally been seen as providing evidence for the sequential adjustment theory of anchoring, according to which time pressure limits the number of adjustments possible, keeping the final response closer to the anchor. As the BAM network proposed in this paper generates sequential adjustment, it is able to provide an explanation for these results as well. The BAM network often does not settle at its stable response in one time step; rather its response nodes activate and deactivate consecutively over time, before stabilising at the final response (as in e.g., Figures 3 and 4). When the decision-maker is faced with time pressure, the network is not allowed to stabilise and the adjustment process generated in this network is curtailed, generating a stronger anchoring effect.

Of course, there are many findings regarding the anchoring effect that the parsimonious structure of the BAM network does not allow it to accommodate. For example, as the model does not distinguish between exemplar and category knowledge, it is unable to capture the findings pertaining to implausible anchors and their associated decision times captured by Mussweiler and Strack (2000). Likewise, the model is largely unable to generate its own anchors (it requires exogenous inputs to activate response nodes), and thus the network cannot account for findings on self vs. experimenter generated anchors documented by Epley and Gilovich (2001). Additionally, although the proposed account of anchoring is able to reconcile many of the differences between the sequential adjustment and selective accessibility theories of anchoring, a third, newer theory of anchoring, based on scale distortion (Frederick & Mochon, 2012), appears to be outside of its scope. Future work on modelling anchoring should attempt to expand on the BAM model and related approaches, to better describe these findings.

In summary, the BAM network provides a simple model for the biased activation theory of anchoring. We have shown that the settling dynamics of this BAM network generate sequential adjustment. Anchors trigger a cascade of activation in the response layer of the BAM network, with nodes in this layer activating and deactivating consecutively. This progression of activation is generally insufficient and final responses depend critically on starting anchor values. By reconciling two contrasting theories within one framework, the BAM network is able to provide a parsimonious explanation for a wide range of findings regarding anchoring and its moderators. These results also imply that other more complex models of bidirectionality should be able to account for anchoring phenomena (as suggested initially in Russo, 2010; see also Glöckner & Englich, 2015).

Choice alternatives

Coherence shifts

A second application of the BAM model pertains to results regarding attributes and alternatives in preferential decision-making tasks. Here, as with evidence–conclusion relationships, the evaluation of choice alternatives is affected by the attributes that are activated, and in turn attribute activation is influenced by the choice alternatives in consideration. This manifestation of bidirectionality has most often been studied in the context of coherence shifts in preferential choice. Decision-makers' evaluations of attributes do not remain constant throughout the decision task. Rather, attribute ratings change over the time course of the task, so that attributes present in preferred choice alternatives are more likely to be rated as important or valuable near the end of the decision, relative to that at the start (Simon et al., 2004, Simon et al., 2008). As an example of this, let us consider Simon et al.'s (2004) experiment in which participants were asked to choose between different jobs. As in Holyoak and Simon's (1999) legal judgement study, participants were also asked to evaluate the specific attributes of the jobs. Simon et al. found that participants' evaluations of attributes changed over the course of the decision, with attributes in the eventually chosen alternative being rated more favourably than they had been rated initially.

In a similar manner to coherence shifts in judgement, preexisting preferences for alternatives alter evaluations of attributes (Russo et al., 1996, 1998. For example, Russo et al. (1996) asked participants to choose between two restaurants. Prior to this, one of the restaurants was experimentally made more preferable to the participants. After this manipulation, but prior to the choice, participants were asked to judge whether different restaurant attributes support the preferred or non-preferred restaurants. As with Simon et al.'s (2004) finding in judgement, Russo et al. found that participants distorted their evaluations of the attributes to cohere with their preferences.

These coherence shifts (and the choices that they generate) are susceptible to various environmental influences. For example, inferior alternatives, that is, alternatives that are not selected in unbiased binary choice tasks, can be chosen by decision-makers if the attributes present in these alternatives are made especially salient at the start of the decision process (Russo et al., 2006). Likewise, the coherence shifts observed in these settings, as well as in related unbiased tasks, can be eradicated by making all of the underlying attributes especially salient for the decision-maker (Carlson & Pearo, 2004). Similar results have also been found in risky choices (Dekay et al., 2009; Russo & Yong, 2011).

We can explain these results within the BAM framework if we assume that choice alternatives are represented by the response layer and attributes are represented by the cue layer. Responses have positive connections with cues if their corresponding alternatives contain the cues' corresponding attributes, and negative relations otherwise. In this setting, the stable states of the network are described by the activation of certain choice alternatives, as well as the activation of their component attributes. With this structure, coherence shifts in choice can be explained by the same properties of the BAM model that are used to explain coherence shifts in judgement. For example, the changes in attribute ratings over time (Simon et al., 2004, Simon et al., 2008) emerge as a product of changes in attribute activation caused by the BAM settling process, in a manner identical to changes in cue activation, and subsequently ratings of cue validity in the experiments of Holyoak and Simon (1999) (and also Glöckner et al., 2010; Kostopoulou et al., 2009). Likewise, the effect of preexisting preferences (Russo et al., 1996, 1998 on attribute ratings and final choices in the BAM network is identical to the response bias effects captured in the work of Simon et al. (2004). Finally, the effects of attribute salience on choice (Carlson & Pearo, 2004; Russo et al., 2006) in the BAM network are identical to the effects of cue salience on final response activation discussed in the anchoring section, with regard to the results of Chapman and Johnson (1999) and Mussweiler et al. (2000). For parsimony, we will not expand on these results here.

Reference point effects

The bidirectional relationship between choice alternatives and component attributes can generate not only coherence shifts in choice, it can also create preference reversals between choice tasks with different salient alternatives.

The dependence of choice on salient alternatives is best illustrated by reference point effects (Tversky & Kahneman, 1991). Decision makers tend to prefer reference points, such as previous endowments or the status quo, over competing alternatives (Knetsch, 1989). They also prefer choice options that dominate reference points and choice options that involve small trade-offs from the reference point, over competing alternatives (Herne, 1998; Trueblood, 2015; Tversky & Kahneman, 1991).

As an example of this phenomenon, let us consider the three experiments outlined in Tversky and Kahneman (1991). The first pertains to the endowment effect as captured by Knetsch (1989). In this experiment, participants are endowed with either a mug or a chocolate bar at the start of the experiment. After a brief period, they are then asked whether they want to trade their endowed object for the competing item. Knetsch (1989) finds that participants typically choose to remain with their endowed item, contradicting rational models of economic choice.

The remaining two experiments in Tversky and Kahneman (1991) are ones they conducted themselves. The experiment for what they label the improvements vs. trade-offs effect involves endowing participants with either one free dinner at a restaurant or one free photo portrait, and then asking them whether they would want to keep their gift or exchange it for two free dinners or a photo portrait with a wallet-size print. Tversky and Kahneman found that participants preferred the novel options that dominated their endowment, with participants endowed with one dinner choosing the two dinner option, and participants endowed with the photo portrait choosing the photo portrait with a wallet-size print option (see also Herne, 1998; Trueblood, 2015).

The third experiment pertains to the advantages and disadvantages effect. In this experiment, Tversky and Kahneman told participants that their present job, which involved very little social contact but also a very short commute, was ending. They now had a choice between a job with little social contact and a short commute, and a job with high social contact and also a long commute. They also ran a mirror variant of this in which the endowed job had very high social contact and a very long commute. Overall, Tversky and Kahneman (1991) found that participants preferred the option that that was more similar to their reference point, so that changing the reference point could reverse participants' choices between the two available options (again, see Herne, 1998; Trueblood, 2015).

Recent works have attempted to explain these effects by arguing that reference points bias attention towards the attributes that they possess (Ashby et al., 2012; Carmon & Ariely, 2000; Johnson et al., 2007; Nayakankuppam & Misra, 2005; Willemsen et al., 2011; see also Russo et al., 1996 discussed above). This bias can increase the importance of these attributes in choice, and can lead to an overall preference for the reference point. Thus, for example, Johnson et al. (2007) find that participants in an endowment effect study who are given mugs are more likely to list mug-related attributes as being important than other participants.

These findings can be captured by the BAM framework if we assume that reference points, like other salient responses, determine the starting state of the network. Again, we will have to assume that alternatives are represented on the response layer, attributes are represented on the cue layer, and positive or negative connections between the nodes in this network are based on the attributes that are contained in the various alternatives. Now consider a setting with two alternatives, and minimal attribute overlap between these alternatives, as shown in the left panel of Figure 5. In this setting, the network will stabilise with the activation of the first response, and cues that support this response, if the alternative corresponding to that response is the reference point. Alternatively, if the alternative corresponding to the second response is the reference point, then the network will stabilise with the activation of the second response, as well as the activation of the cues that support this response. This will generate the endowment effect due to the same mechanism responsible for response biases in the judgement tasks discussed above, and the cue overlap relationships that characterise these response biases (again, discussed above) will determine whether or not the endowment effect ultimately emerges.

These insights can also, however, be used to explain the two other findings regarding reference-dependent choice. Again, the first of these pertains to the improvements vs. trade-offs effect, which refers to the preference for alternatives that dominate the reference point, that is, alternatives that contain all of the attributes in the reference point, as well as some attributes not contained in the reference point (Herne, 1998; Trueblood, 2015; Tversky & Kahneman, 1991). BAM is able to explain this result if we assume that the reference point determines the starting state of the network. Consider, for example, the middle panel of Figure 5, which is identical to the left panel except for the presence of response 3, a dominated response option. If the network begins with the activation of this dominated response option (when it is the reference point), then the cues that support this option, cues in the set C_3, will be activated alongside this option in the first time period. Subsequently, response 2, which is supported by the majority of the cues in this set will activate in the second time period. Response 1, corresponding to the alternative that does not dominate the reference point, will remain deactivated throughout this process. More formally, if we have $C_3 \subset C_2$ and $C_3 \cap C_1 = \emptyset$, as well as $|C_2 \cap C^{c}{}_{1}| \geq |C_2 \cap C_1|$, then activating response 3 at the start of the decision will necessarily lead to the activation of response 2, but not response 1, at the end of the decision.

The second finding pertains to the advantages vs. disadvantages effect, which refers to the preference for alternatives that involve few trade-offs from the reference point. Consider the right panel of Figure 5, which is identical to the middle panel, except for the fact that response 3 shares a lot of attributes with response 2, but not with response 1 (Trueblood, 2015; Tversky & Kahneman, 1991). If response 3 is the reference point, then response 2 can be seen as involving few trade-offs from the reference point. In this setting, BAM predicts that response 2 will also be activated once the network stabilises. This is because the large cue overlap between responses 3 and 2 leads to a spread of activation from response 3 to response 2. Because there is little overlap between response 2 and response 1 (and no overlap between response 3 and response 1), activation will not spread to response 1. More formally, if we have $|C_3 \cap C_2| > |C_3 \cap C^c_2|$, as well as $|C_3 \cap C^c_1| \geq |C_3 \cap C_1|$ and $|C_2 \cap C^c_1| \geq |C_2 \cap C_1|$, then activating response 3 at the start of the decision will necessarily lead to the activation of response 2, but not response 1, at the end of the decision. Note that this is the same mechanism responsible for the activation-based sequential adjustment biases discussed in the anchoring task. Also note that the biases discussed above would emerge even if the reference point was not part of the choice set: All that is needed is a starting point bias that influences attribute activation.

To summarise, this section has used the BAM network to study the relationship between bidirectionality and various reference point effects, including the endowment effect, the improvements vs. trade-offs effect and the advantages vs. disadvantages effect. It has shown how the cue overlap and similarity-based spread of activation in the BAM network can account for these findings if it is assumed that reference points determine the starting activation of the network (see also Bhatia, 2013 for a related approach). As the BAM network is a simplified variant of other existing bidirectional models, our results suggest that these models should also be able to capture the findings discussed in this section.

Novel predictions

Perhaps, the most important property of the BAM network pertains to cue overlap and similarity-based spreading activation. As discussed above, activating one response activates another only if the cues supporting the second response overlap also support the first. Subsequently, increasing the extent of cue overlap between two responses increases the probability that activation will spread from one response to the other. This can be rigorously tested. In judgement without numerical hypotheses, this would imply that making one conclusion (such as a verdict in a court case) particularly salient to the decision-maker at the start of the judgement task (as in e.g., Simon et al., 2004) increases the probability that a second conclusion will be chosen as a response in this task in proportion to the similarity between the two conclusions. In tasks with numerical responses, this would imply that the spread of activation from one number to another, and subsequently the strength of the anchoring effect, would depend on the degree to which cues support multiple neighbouring responses. Likewise, in choice tasks, cue overlap and similarity-based spreading activation predicts that instances of reference dependence, such as the endowment effect, may weaken with the proximity of a competitor to the reference point.

The above sections also show that this similarity-based spreading activation is asymmetric. Activating one response may activate another, but not vice versa. This asymmetry again stems from cue overlap relationships, and this too can be tested in judgement and choice using anchors, reference points, and other salient responses. Thus, for example, endowing a decision-maker with a choice option whose attributes also, on average, support its competitor, but not vice versa, may lead to a weak endowment effect. But doing the opposite may lead to a fairly strong endowment effect.

Similarity can also enable us to develop predictions pertaining to decision time. Recall that the time that the network takes to stabilise is assumed to be a proxy for decision time. Thus, decisions in which the activation is most likely to continue spreading for some time should be the decisions that take the longest. These decisions are ones involving multiple similar conclusions or choice alternatives, in which there is considerable cue overlap for the responses. In contrast, decisions involving many dissimilar conclusions or choice alternatives should be quicker.

A fourth set of predictions emerge from our characterisation of the network's stable state. As discussed above, not only is the network guaranteed to stabilise, but it is also guaranteed to stabilise on responses that are supported by more than half of the activated cues. It is possible to test this in simple settings by asking decision-makers to make judgement and choices, and then to rate various pieces of evidence or various attributes pertaining to the decision. According to the proposed model, more than half of the cues rated positively by the decision-makers should support their selected responses (and the opposite for cues rated negatively by decision-makers).

The predictions described here are only a small portion of those generated by the model. By clearly specifying how cue–response relationships and starting points drive the behaviour of bidirectional processes, the BAM network is able to make a vast range of analytically grounded quantitative predictions, which can be fit with sufficient data. Examining these predictions, either qualitatively or quantitatively, should be the focus of future work. Additionally, note that these predictions are not unique to BAM; they should emerge from more complex bidirectional models, such as Co3 and PCS-DM (Glöckner & Betsch, 2008; Holyoak & Simon, 1999) as well. However, it is BAM's analytical tractability that allows us to derive these predictions, and to clearly understand the ways in which they relate to bidirectionality and to existing behavioural effects attributed to bidirectionality.

Discussion and conclusion

We have attempted to understand the properties of high-level bidirectional cognition by studying the BAM proposed by Kosko (1988). This network involves simplified assumptions regarding activation states and connection weights, and is subsequently mathematically tractable. We have shown that this model can be used to organise a range of findings in judgement and decision-making research, including coherence shifts, anchoring, and reference dependence, and that the application of this model to judgement and decision-making generates a number of novel testable predictions. Both these novel predictions and the existing behavioural effects captured by BAM should extend to existing models like Co3 and PCS-DM (Glöckner & Betsch, 2008; Holyoak & Simon, 1999), with which BAM shares its main properties.

BAM is a toy model, and for this reason there are some limitations to using the BAM network to study judgement and decision-making. The properties that make BAM mathematically tractable also prevent it from being able to capture more complex phenomena. For example, binary activations imply that the network cannot, in its current form, be used to give continuous responses, such as responses corresponding to strength of belief in judgement, or to strength of preference in choice. Likewise, unlike human judgement and decision-making, the model is deterministic, and gives the same responses across trials if its parameters and starting states are kept constant. Finally, binary cue—response connection weights make it impossible for the model to represent continuous relationships between cues and responses, as is the case in many relevant decision-making domains.

The goal of this paper is not, however, to provide a complete and comprehensive model of bidirectional decision-making (for this purpose, models such as Co3 and PCS-DM, which permit continuous activation states and continuous connection weights, do a suitable job), but rather to use the BAM network to formally characterise the complex computations and behaviours that existing bidirectional models entail. Doing so not only improves our understanding of these models, and of bidirectionality more generally, but also provides a number novel insights regarding the relationships between findings in various diverse domains. For example, as cue overlap determines assessments of similarity as well as the spread of activation from a salient response (such as that provided by an anchor) to another, the anchoring effect can be reinterpreted as a type of similarity-based starting point bias, and vice versa. With the same logic, anchoring and similarity processing can be seen as different interpretations of reference dependence in preferential choice, and response biases in judgement, and so on. All of these different phenomena are, in turn, products of the natural dynamics of coherence maximisation that are fundamental to bidirectional processing.

It would be valuable to extend the results of this paper to other domains. For example, when outlining the relationship between spreading activation and asymmetric similarity assessments, this paper suggested that the nodes in the two layers of the BAM network could correspond to categories and features. One could take this interpretation further, and use the BAM network to characterise not only similarity judgement between categories, but also phenomena as diverse as semantic priming, feature induction, and category learning.

Another more complex application of the BAM network involves game theoretic decision-making, as outlined by Bhatia and Golman (2014). Bhatia and Golman (2014) show that the settling dynamics of the BAM network can, amongst other things, recover fundamental solution concepts like pure-strategy Nash equilibria. Further work should attempt to apply the BAM model, and its insights, to study how humans solve strategic games. It seems, after all, that behavioural findings on strategic choice, category and feature processing, reference dependence, and anchoring may not be entirely independent. Rather, they could all be highly interrelated, and highly complementary, implications of bidirectional processing in human cognition.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Sudeep Bhatia http://orcid.org/0000-0001-6068-684X

References

Ashby, N. J. S., Dickert, S., & Glöckner, A. (2012). Focusing on what you own: Biased information uptake due to ownership. *Judgment and Decision Making*, *7*, 254–267.

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, *120*, 522–539.

Bhatia, S., & Golman, R. (2014). A recurrent neural network for game theoretic decision making. *Proceedings of the 36th annual conference of the Cognitive Science Society*. Quebec City, Canada.

Bond, S. D., Carlson, K. A., Meloy, M. G., Russo, J. E., & Tanner, R. J. (2007). Information distortion in the evaluation of a single option. *Organizational Behavior and Human Decision Processes*, *102*, 240–254.

Brownstein, A. L. (2003). Biased predecision processing. *Psychological Bulletin*, *129*, 545–564.

Carlson, K. A., & Pearo, L. K. (2004). Limiting predecisional distortion by prior valuation of attribute components. *Organizational Behavior and Human Decision Processes*, *94*, 48–59.

Carmon, Z., & Ariely, D. (2000). Focusing on the forgone: How value can appear so different to buyers and sellers. *The Journal of Consumer Research*, *27*, 360–370.

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making*, *7*, 223–242.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, *79*, 115–153.

DeKay, M. L., Patiño-Echeverri, D., & Fischbeck, P. S. (2009). Distortion of probability and outcome information in risky decisions. *Organizational Behavior and Human Decision Processes*, *109*, 79–92.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*, 391–396.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*, 124.

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*, 215–228.

Glöckner, A., Betsch, T., & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making*, *23*, 439–462.

Glöckner, A., & Betsch, T. (2012). Decisions beyond boundaries: When more information is processed faster than less. *Acta Psychologica, 139*, 532–542.

Glöckner, A., & Englich, B. (2015). When relevance matters: Anchoring effects can be larger for relevant than for irrelevant anchors. *Social Psychology*, *46*, 4–12.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, *133*, 641–666.

Glöckner, A., & Hodges, S. D. (2011). Parallel constraint satisfaction in memory based decisions. *Experimental Psychology*, *58*, 180–195.

Guo, F. Y., & Holyoak, K. J. (2002). Understanding similarity in choice behavior: A connectionist model. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society* (pp. 393–398). Fairfax, Virginia.

Herne, K. (1998). Testing the reference-dependent model: An experiment on asymmetrically dominated reference points. *Journal of Behavioral Decision Making, 11*, 181–192.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*, 3–35.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295–355.

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review*, *112*, 841.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 461–474.

Knetsch, J. L. (1989). The endowment effect and evidence of nonreversible indifference curves. *The American Economic Review, 79*, 1277–1284.

Kosko, B. (1987). Adaptive bidirectional associative memories. *Applied Optics*, *26*, 4947–4960.

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man and Cybernetics*, *18*, 49–60.

Kostopoulou, O., Mousoulis, C., & Delaney, B. C. (2009). Information search and information distortion in the diagnosis of an ambiguous presentation. *Judgment and Decision Making*, *4*, 408–418.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284–293.

Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: The ACS (Attitudes as Constraint Satisfaction) model. *Psychological Review*, *115*, 733–745.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*, 136–164.

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology*, *78*, 1038–1046.

Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes*, *86*, 234–255.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*, 1142–1150.

Nayakankuppam, D., & Mishra, H. (2005). The endowment effect: Rose-tinted and dark-tinted glasses. *Journal of Consumer Research*, *32*, 390–395.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*, 429–439.

Reitsma-van Rooijen, M., & L Daamen, D. D. (2006). Subliminal anchoring: The effects of subliminally presented numbers on probability estimates. *Journal of Experimental Social Psychology*, *42*, 380–387.

Russo, J. E. (2010). Understanding the effect of a numerical anchor. *Journal of Consumer Psychology*, *20*, 25–27.

Russo, J. E., Carlson, K. A., & Meloy, M. G. (2006). Choosing an inferior alternative. *Psychological Science*, *17*, 899–904.

Russo, J. E., Carlson, K. A., Meloy, M. G., & Yong, K. (2008). The goal of consistency as a cause of information distortion. *Journal of Experimental Psychology: General*, *137*, 456–469.

Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes*, *66*, 102–110.

Russo, J. E., Meloy, M. G., & Medvec, V. H. (1998). Predecisional distortion of product information. *Journal of Marketing Research*, *35*, 438–452.

Russo, J. E., & Yong, K. (2011). The distortion of information to support an emerging evaluation of risk. *Journal of Econometrics*, *162*, 132–139.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, *103*, 219.

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review*, *6*, 283–294.

Simon, D., Krawczyk, D. C., Bleicher, A., & Holyoak, K. J. (2008). The transience of constructed preferences. *Journal of Behavioral Decision Making*, *21*, 1–14.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004). Construction of preferences by constraint satisfaction. *Psychological Science*, *15*, 331–336.

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1250–1260.

Simon, D., Snow, C. J., & Read, S. J. (2004). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology*, *86*, 814.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency. *Journal of Social Issues*, *4*, 147–165.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*, 437–445.

Switzer, F. S., & Sniezek J. A.. (1991). Judgment processes in motivation: Anchoring and adjustment effects on judgment and behavior. *Organizational Behavior and Human Decision Processes*, *49*, 208–229.

Trueblood, J. S. (2015). Reference point effects in riskless choice without loss aversion. *Decision*, *2*, 13–20.

Thagard, P. (1989). Explanatory coherence. *Behavioral and brain sciences*, *12*, 435–467.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*, 1039–1061.

Whyte, G., & Sebenius, J. K. (1997). The effect of multiple anchors on anchoring in individual and group judgment. *Organizational Behavior and Human Decision Processes*, *69*, 74–85.

Willemsen, M. C., Böckenholt, U., & Johnson, E. J. (2011). Choice by value encoding and value construction: Processes of loss aversion. *Journal of Experimental Psychology: General*, *140*, 303–324.

Appendices

Appendix 1. Details of model properties

Spreading activation and cue overlap

One of the most important properties of bidirectional networks (and related neural and associative networks) pertains to the spread of activation from one response node to another. Particularly, as activated responses determine the activation of cues, which in turn determine the further activation of various responses, activating one response can lead to the activation of another. This happens despite the fact that response nodes do not directly connect with each other. Understanding the spread of activation across responses, and the state transition rules that characterise this spreading activation, is necessary for fully characterising BAM's behaviour.

For now, we will consider a setting with only two responses, 1 and 2, and some M number of cues. Assume that at time $t = k$, response 1 is activated, and response 2 is deactivated. This could be because response 1 is especially salient at the start of the judgement or decision-making task (and $t = 1$), or because the network has evolved to this state over the time course of the deliberation process. Due to this activation pattern, cues in the set C_1 are also activated at $t = k$. These are cues that receive net-positive inputs from the response layer and thus have inputs that are greater than 0, the threshold for activation. Intuitively, the decision-maker focuses on all the cues that are positively related to the activated response and suppresses all cues that are negatively related to the activated response. Note that cues not in the set C_1 are not activated, as these cues receive inhibitory inputs from response 1, leading to total inputs less than or equal to 0.

Once these cue nodes are activated, the activation pattern in the response layer can change. This change depends critically on the structure of cue overlap between the various nodes in the response layer. Particularly, at $t = k + 1$, responses which are positively related to the majority of the cues in C_1 turn on. These include not only the initially activated response 1 (which is supported by all cues in C_1), but also other novel response options. Responses that are opposed by the majority of the cues in C_1 stay off. For the two response setting explored here, we can write the condition for activation at time $t = k + 1$, as follows: if response 1 is the only activated response at $t = k$, then response 2 activates at $t = k + 1$ if and only if BAM's encoded memory is such that $|C_1 \cap C_2| > |C_1 \cap C^c_2|$. Intuitively, if the cues that are positively related to an activated response also support a second response, then the second response will activate. Can the new pattern of activation in the response layer eventually deactivate response 1? If the cues that are activated at $t = k + 1$ are, on average, negatively related to response 1, then response 1 will indeed turn off at time $t = k + 2$. More formally, if the activation of response 1 leads to the activation of response 2 at time $t = k + 1$, then $C_1 \cup C_2$ is the set of cues that are also active at $t = k + 1$. Subsequently, response 1 will turn off at $t = k + 2$ if the cues that are negatively related to response 1, and subsequently inhibit response 1, make up the majority of this set, that is, if $|C^c_1 \cap [C_1 \cup C_2]| \geq |C_1 \cap [C_1 \cup C_2]|$, which simplifies to $|C^c_1 \cap C_2| \geq |C_1|$.

Figure A1 displays three examples of cue overlap that can generate the three types of dynamics observed in a two-response setting, with response 1 activated and response 2 deactivated at time $t = k$. In the left panel, response 1 does not overlap sufficiently with response 2, for it to activate response 2 at $t = k + 1$ (i.e., $|C_1 \cap C^c_2| \geq |C_1 \cap C_2|$). In the middle panel, the two responses overlap just enough (i.e $|C_1 \cap C_2| > |C_1 \cap C^c_2|$ and $|C_1| > |C^c_1 \cap C_2|$) for response 2 to activate at $t = k + 1$, and for both responses to be remain activated at $t = k + 2$. In the right panel, response 1 overlaps sufficiently with response 2, but not vice versa ($|C^c_1 \cap C_2| \geq |C_1|$) so that response 2 activates

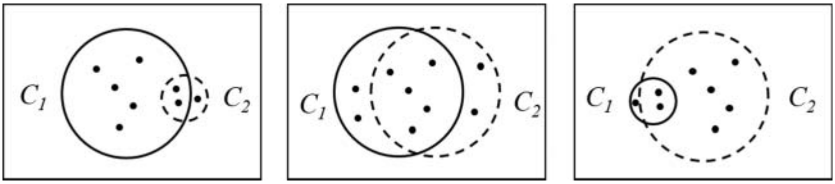


Figure A1. Three examples of cue overlap. Here, the number of black dots within the circle corresponds to the number of cues supporting response i, $|C_i|$. In the first panel, activating response 1 does not activate response 2. In the middle panel, activating response 1 activates response 2, and both responses remain activated. In the right panel, activating response 1 activates response 2 which in turn deactivates response 1.

at $t = k + 1$, but then causes response 1 to deactivate at $t = k + 2$. The key insight for these three settings is the following: activation spreads from one response to another if the cues that are positively related to the first response are also positively related to the other. This is because cues that are positively related to an activated response are themselves activated with that response, and because responses that are positively related to activated cues are themselves (in the subsequent time period) activated with those cues. If the cues that are positively related to the second response are not also positively related to the first, then the second response's activation can deactivate the first.

Similarity

The above section highlights the relationships between spreading activation and cue overlap: activation spreads from one response to another if they overlap sufficiently. Now, similarity assessments are often based on cue (or feature) overlap, and this implies that similarity may be closely related to bidirectional processing. With this in mind, it may be possible to use the behaviour of the BAM network, which involves spreading activation between similar response nodes, to capture results regarding similarity assessments in other domains.

Consider, for example, the well-known asymmetry of feature-based similarity assessment. More people agree with the statement "North Korea is similar to China" than with the statement "China is similar to North Korea". Phenomena such as this are typically explained by the dependence of feature attention on the focal category, as in Tversky's (1977) contrast model. In this model, the similarity index is an increasing function of the number of overlapping features of the two categories, and a decreasing function of the number of non-overlapping features of the two categories. The observed asymmetries in similarity assessment can be explained if decision-makers are more likely to attend to the features contained in the focal category that are absent in the non-focal category, than the reverse.

The mechanisms in Tversky's model are bidirectional, with focal categories determining feature attention, and salient features in turn determining assessments of category similarity. Can we, with the assumption that BAM's cue layer represents features and its response layer represents categories, use BAM to capture this important finding on similarity assessment? Let us assume that the focal category is activated at the start of the similarity assessment, and that two categories are judged as similar if activation spreads from the focal category to the non-focal category, and not similar otherwise. In a two-response setting, with response 1 corresponding to the focal category and response 2 corresponding to the non-focal category, this happens only if $|C_1 \cap C_2| > |C_1 \cap C^c{}_2|$. For assessments of similarity when response 2 is the focal category, we require an alternate set of conditions, that is, we need $|C_2 \cap C_1| > |C_2 \cap C^c{}_1|$. It is possible for this second condition to be violated even if the first condition is not, and indeed the settings in which this happens are precisely those in which Tversky's model yields asymmetries (i.e., when there is asymmetric cue overlap between the two categories). The right panel of Figure A1 provides an example of cue overlap which illustrates this insight. Here, responses 1 and 2 are such that we have $|C_1 \cap C_2| > |C_1 \cap C^c{}_2|$, but we do not have $|C_2 \cap C_1| > |C_2 \cap C^c{}_1|$. Subsequently, if the network begins with response 1 activated, activation spreads to response 2 as well, generating a positive assessment of similarity. If, however, the network begins with response 2 activated, then activation does not spread to response 1, and the network stabilises with only response 2 activated. In this case, the two categories will not be judged as similar.

This analysis shows that the properties of the BAM network that determine the spread of activation across responses place more weight on the unique features of the focal category than they do on the unique features of the non-focal category. This can lead to asymmetric similarity assessments, allowing BAM to capture some of the findings explained by Tversky's model. It is, however, possible to make a claim that is stronger than this. Particularly, if we simplify Tversky's (1977) model so that similarity is a binary relation, with one category being similar to another if the similarity index for the two categories exceeds a threshold of 0, then we can prove that the BAM network exactly instantiates Tversky's model with certain parametric restrictions. In other words, the similarity assessments generated by BAM are identical to those generated by a parsimonious version of Tversky's model, and BAM can subsequently be seen as representing a special case of Tversky's model.

To see this, note that we can represent Tversky's (1977) contrast model using the following equation:

$$S(r_1, r_2) = \theta \cdot |C_1 \cap C_2| - \alpha \cdot |C_1 \cap C_2^c| - \beta \cdot |C_2 \cap C_1^c|.$$

Here, r_1 is the focal category, and θ, α, and β are non-negative parameters. $S(r_1, r_2)$ is a similarity interval scale. If $\alpha = \beta = 0$, then we have $S(r_1, r_2) = S(r_2, r_1)$, leading to symmetric similarity judgements.

We have assumed that r_1, the focal category, is judged to be similar to r_2 if activating r_1 also activates r_2. This happens if $|C_1 \cap C_2| > |C_1 \cap C^c{}_2|$, which corresponds to the contrast model with $\theta = \alpha$, $\beta = 0$, and a binary similarity threshold such that r_1 is similar to r_2 if $S(r_1, r_2) > 0$.

Coherence and stability

In the above section, we have examined the conditions under which activation spreads in the BAM network. Could these conditions generate spreading activation that continues indefinitely, with certain response nodes activating sequentially or chaotically over time? If this is the case, the network will not stabilise and the decision-maker will not be able to provide a response.

This type of behaviour is not possible for the two-response setting described above. A response node, activated by itself, cannot turn off (or subsequently turn on again), as it necessarily has a positive relationship with the cues that it activates, a relationship that keeps it on as long as no other responses are activated. If activation spreads to a neighbouring response, the second response can either turn off the first response or let it stay on, but it cannot create a pattern of unstable oscillatory activation by first turning it off and then turning it back on again. This would require both $|C^c{}_1 \cap C_2| \geq |C_1|$ and $|C_2 \cap C_1| > |C_2 \cap C^c{}_1|$, which is not possible.

Indeed, based on the analysis in Kosko (1988), we can prove that any BAM network, with any number of nodes, with any encoded memories, starting at any point, with any (constant) exogenous inputs will necessarily stabilise, and do so in a finite number of time steps. The intuition for this claim is the following. At any time t, the activation state of the BAM network can be described by the following coherence level (which is identical to that assumed in Thagard, 1989): $\mathrm{Coh}[\mathbf{r}(t), \mathbf{c}(t)] = \sum_{i=1}^{N} \sum_{j=1}^{M} r_i(t) \cdot w_{ij} \cdot c_j(t)$. If a particular state is such that the activated cues do not support the activated responses, then that state has a low level of coherence. Likewise, if a state is such that all the activated cues support all the activated responses, then that state has a high level of coherence. It is possible to show that the pattern of spreading activation, discussed in the section above, necessarily increases the coherence level in the network. As there are a finite number of nodes in the network, and each node can be either on or off, the network can only be in a finite number of states. Subsequently, as changes in activation states increase coherence, after a finite number of time steps, the network should reach a state of maximal (local) coherence, a state from which further changes are not possible. This state will be stable.

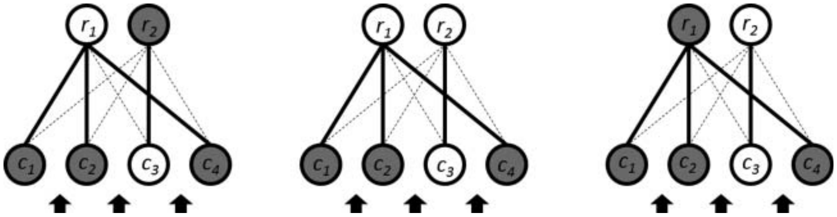


Figure A2. Three examples of cue and response activation. Here, response 1 is supported by cues 1, 2, and 4 (indicated by solid lines), but not cue 3 (indicated by a dotted line). In contrast, response 2 is supported only by cue 3, and not by the other cues. The activation states in the first two panels are unstable, and will change to a nearby state of increased coherence. Only the activation state shown in the right panel is stable.

We can gain some additional insights on the relationship between stability and coherence by examining the types of states that are stable. Consider, for example, an incoherent state in which an activated response is opposed by the majority of activated cues. An example of this is illustrated in the left panel of Figure A2. This state can never be stable. As more than half of the activated cues do not support the activated response, the response will get net negative inputs and will deactivate in the subsequent time period. Likewise, an incoherent state in which an inactivated response is supported by the majority of activated cue (as in the middle panel of Figure A2) is unstable. Intuitively, decision-makers will change their mind if the response they are considering is not supported by the majority of salient cues, or if the majority of salient cues support an unconsidered response. Indeed, it is easy to show that a response can only be activated in a stable state if it is supported by at least half of the cues that are activated in that state, and a set of cues can only be activated in a stable state if the responses that they support are activated in that state (as in the right panel of Figure A2). Decision-makers will only select responses that are coherent with activated cues.

Another property of coherence in the BAM network involves similarity. BAM's dynamics operate primarily through similarity, and it is these similarity-based dynamics that lead to increased coherence in the network. Intuitively, a coherent decision-maker who believes one conclusion based on a certain set of evidence should also believe other conclusions that are implied by this evidence. Likewise a coherent decision-maker who prefers one alternative after considering a set of attributes should also prefer other alternatives that share these attributes. More generally, considering one response option without considering other responses that are similar to it is an incoherent state, one that is naturally avoided by the dynamics of the BAM network. Indeed, a state is stable only if responses that are similar enough to each other are activated concurrently.

Stability, coherence, and similarity are, thus, closely related aspects of bidirectionality. Similarity leads to increasing coherence which then guarantees stability. This result thus provides valuable insights about how these models are able to maximise coherence. This result also has practical benefits. By guaranteeing a final, unique, set of activated responses, it ensures that there is always a solution to whatever psychological task the BAM model is applied to.

Exogenous inputs

Thus far, we have derived results regarding spreading activation, its dependence on cue overlap and response similarity, and its implications for network stability. These results, however, pertain only to the encoded memories of the BAM network, which are not enough, by themselves, to determine the network's behaviour. To fully characterise BAM's behaviour, we need to understand the effect of exogenous inputs into the network on the spread of activation in the network.

As specified earlier, we assume that temporary exogenous inputs $I^r_i(t) = 1$ to the response layer at the start of the decision process ($t = 0$) determine the starting activation states for the response nodes. Responses that are particularly salient at the start of the decision process receive positive inputs, whereas those that are not particularly salient do not receive any inputs. In this sense, these inputs serve as a form of response bias, which determines the initial conditions of the network. We assume that there are no inputs to the response layer after $t = 0$.

Now, in many situations, the decision-maker has no reason to favour any one response, and, thus, we can assume that the network in these settings has $I^r_i(0) = 0$ and begins with all the response nodes turned off. This response activation state sends no feedback into the cue layer, and the cues in the set of relevant cues C all activate at the first time period. In this setting, the responses that are supported by the majority of cues in C are the only ones that receive net positive inputs in the subsequent time period, and are the only ones that are activated. In later sections, we shall see that these majority responses have a special normative status, and we will examine the settings in which they remain activated in more detail.

Many decision tasks do involve a response bias, that is a focal response that is activated at the start of the decision. In these settings, $I^r_i(0) = 1$ activates the focal response and activation spreads from this focal response to other responses, based on similarity, as described in the preceding sections. Changing the focal response affects the trajectory of spreading activation, and subsequently the final stable response.

Are there settings in which these focal responses are guaranteed to remain stable? Yes. If there is little cue overlap between the focal response and other responses, then the focal response and the cues that support the focal response are the only nodes that are activated when the network stabilises. Formally, this happens whenever the majority of the cues that support the the focal response do not support any other response, that is, when the conditions for spreading activation described in the above sections do not hold for any response. Intuitively, this happens because activation does not spread when responses are dissimilar. Without spreading activation, the exogenous inputs that determine initial conditions fully characterise the final stable state of the network. Or, put another way, without spreading activation, the initial state is the one with the maximal local coherence.

Recall, we also assume another set of exogenous inputs into the network. These inputs, $I^c_j(t) = 1$, affect the cue nodes, and are persistent over the time course of the decision. Although they are not a feature of Kosko's original BAM model, they are particularly desirable for the psychological tasks that we are modelling. These inputs ensure that cue nodes are activated even when none of the response nodes are active, and that the judgement or decision process can begin in the absence of a response bias, and continue even if all responses extinguish. They also allow us to model the effect of task-related determinants of cue salience. In some of the experiments we are going to model, certain cues are made more salient than others, and cue salience is varied across the time course of the decision. In these settings, we will assume that I^c_j can vary across cues and across time.

The exogenous inputs characterised by $I^c_j(t)$ also have another useful property: They allow the decision-maker to exert control on the behaviour of the BAM network. Inputs $I^c_j(t) > 1$ lead to a reduction in the relative strength of the feedback from the response layer to the cue layer. If the strength of the inputs is higher than the total number of response nodes (i.e $I^c_j(t) > N$), then the network behaves unidirectionally. That is, all cues in C remain activated throughout the decision process, even if they receive inhibitory inputs from some or all of the responses. Subsequently, the response that is supported by the majority of the cues in C will turn on and then remain activated throughout the remainder of the decision process. Again, the majority responses that these settings activate have special normative relevance. We shall explore this in more detail in the next section.

Correct responses

Before applying the BAM network to the biases studied in judgement and decision-making, it is useful to consider the settings in which the network is able to give correct responses. We would not, after all, expect decision-makers to deliberate in a bidirectional manner, if this type of processing is completely maladaptive. As a result, examining BAM's ability to recover the correct response is in many ways a necessary first step to testing its descriptive power.

In this paper, we will consider a response to be correct if there are more cues that support it than there are that oppose it. More formally, a response i^* is correct if we have $|C_{i^*}| > |C^c_{i^*}|$. While it is possible for many responses to be correct under this definition, there is one important setting where there is guaranteed to be a maximum of one correct response. This is the setting in which there is no overlap in cue support across different responses, as with mutually exclusive conclusions. If each cue supports only one response, then a correct response, if it exists, is guaranteed to be unique.

In this setting, the BAM network will necessarily stabilise with the activation of the correct response (and only the correct response) if there are no exogenous inputs into the response layer, and the network begins without a response bias. All the cues that support the correct response will also be activated in this stable state. Recall that when every node in the response layer is deactivated at the start of the decision, then the exogenous inputs into the cue layer activate all of the cue nodes. These nodes subsequently send net positive inputs into the correct response (which a majority of them support) and net negative inputs into all the other incorrect responses (which a majority of them oppose). As there is no cue overlap, only the correct response turns on. An example of this process is provided in Figure A3.

We can interpret this pattern of activation in terms of coherence. Particularly, in settings where there is no cue overlap, the only stable states in the network correspond to the co-activation of individual responses with their various supporting cues. Thus, in a setting with N responses, each of which are associated with a set of non-overlapping cues C_i such that $|C_i| > 0$, there are N stable states in the network, with each stable state involving $r_i(t) = 1$, $c_j(t) = 1$ for cue $j \in C_i$, and $r_k(t) = 0$ and $c_k(t) = 0$ for all other responses and cues. The coherence of each of these stable states is equal to $\mathrm{Coh}_i = |C_i|$. Subsequently, the state with the correct response activated is the state with the globally maximal coherence.

Note that there is one setting in which the stable state necessarily involves the activation of correct responses, even in the presence of cue overlap and exogenous inputs to responses. This is the setting in which there are high exogenous inputs, $I^c{}_j(t) > N$. As described above, the BAM network in this setting generates unidirectional processing. In particular, exogenous inputs to the cue nodes are higher than any negative feedback these nodes can obtain from the response layer, and the cue nodes remain on through the decision process. This happens even when the network begins with some response nodes activated. When all of the cue nodes are on, all correct responses (there may be multiple when responses can overlap on their cues) receive net positive inputs and remain on. Similarly, all incorrect responses receive negative inputs and remain off.

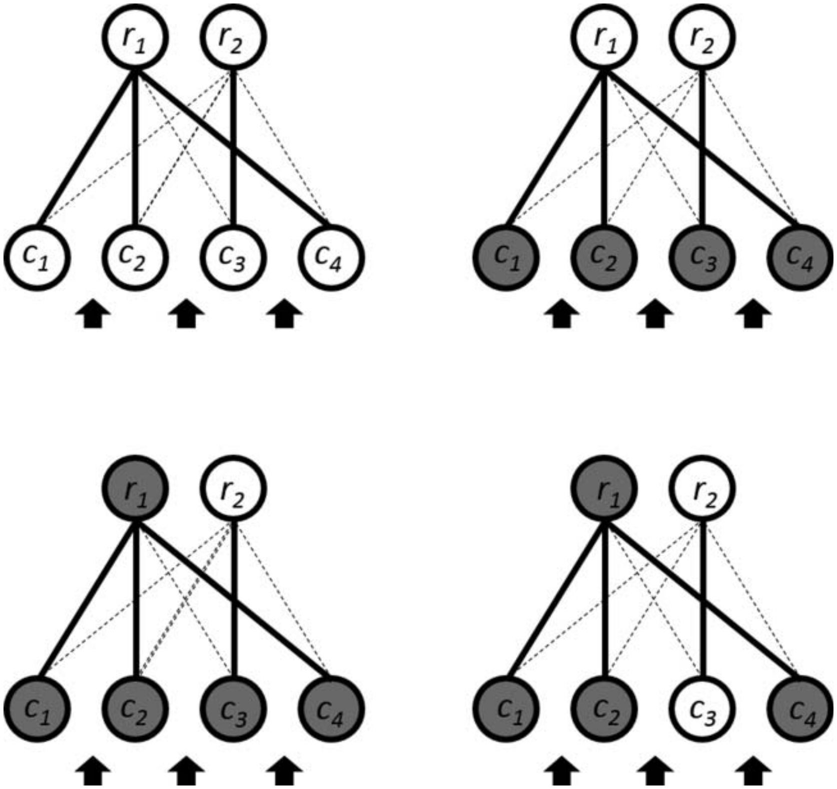


Figure A3. The spread of activation when the network begins without any response bias. Here, all nodes begin deactivated (top left panel). The inputs to the cue layer subsequently activate all cue nodes (top right panel), which then activate the correct response (bottom left panel). After activating, the correct response deactivates cues that do not support it, leading to network stability (bottom right panel).

Appendix 2. Anchoring

Here, we shall show that BAM networks with connected memory structures satisfy contiguous activation and sequential transition. Let us define $C(t)$ to be the set of cues activated at t and $R(t)$ to be the set of responses activated at t. For simplicity, we will refer to response i as r_i and cue j as c_j. Now, consider the following propositions:

Proposition 1a: If a contiguous interval of responses, $r_i, r_{i+1}, \ldots r_k$ is activated at t (and all other responses are deactivated at t), and for $l > k$, r_l is activated at $t + 1$, then it is the case that $r_k, r_{k+1} \ldots r_{1-1}$ are activated at $t + 1$.

Proof: $c_j \in C(t)$ implies $c_j \in C_i \cup C_{i+1} \ldots \cup C_k$. Since $r_l \in R(t+1)$, we have $|C_l \cap C(t)| > |C(t)|/2$. Connectedness implies that if $c_j \in C_i \cup C_{i+1} \ldots \cup C_k$ and $c_j \in C_l$ then $c_j \in C_{l'}$ for $l > l' \geq k$. Hence, if $|C_l \cap C(t)| > |C(t)|/2$, we also have

$|C_{l'} \cap C(t)| > |C(t)|/2$ for all $l > l' \geq k$, which means that $r_l \in R(t+1)$ implies $r_{l'} \in R(t+1)$ for $l > l' \geq k$.

Proposition 1b: If a contiguous interval of responses, $r_i, r_{i+1}, \ldots r_k$ is activated at t (and all other responses are deactivated at t), and for $l < i$, r_l is activated at $t + 1$, then it is the case that $r_{l+1}, r_{l+2}, \ldots r_i$ are activated at $t + 1$.

Proof: The proof for this is identical to that for Proposition 1a.

Proposition 2: If a contiguous interval of responses, $r_i, r_{i+1}, \ldots r_k$ is activated at t (and all other responses are deactivated at t), then for any p and q with $k > p > q > i$, if r_q and r_p are activated at $t + 1$, then so is any r_l for $p > l > q$.

Proof: $c_j \in C(t)$ implies $|R_j \cap R(t)| \geq |R(t)|/2$. As R_j is contiguous (by connectedness), and $R(t)$ is contiguous, and $R_j \cap R(t)$ is also contiguous. Hence, if $c_j \in C(t)$, it supports at least $|R(t)|/2 = (k - i + 1)/2$ contiguous responses in $R(t)$. Assume that $q < (k + i)/2$. If $c_j \in C_q \cap C(t)$, then as c_j supports at least $(k - i + 1)/2$ neighbouring responses in $R(t)$, we must also have $c_j \in C_{q+1}$. Hence, if $|C_q \cap C(t)| > |C(t)|/2$, as is implied by $r_q \in R(t + 1)$, then we have $|C_{q+1} \cap C(t)| > |C(t)|/2$, which implies that $r_{q+1} \in R(t + 1)$. Now, we can use this method again to show that $r_{q+2} \in R(t + 1)$, and keep iterating it to show that $r_l \in R(t + 1)$ for all $(k + i)/2 \geq l \geq q$. Now if $(k + i)/2 \geq p$, then our proof is done. If not, then note that we can use the same logic as above to show that $r_l \in R(t + 1)$ for $p \geq l \geq (k + i)/2$. This then gives us our result.

Now, Propositions 1a and 1b show that, if a contiguous interval of responses is activated at time t, then a response that does not neighbour this contiguous interval cannot be activated at $t + 1$ without activating all intermediate responses. Proposition 2 shows that if a contiguous interval of responses is activated at time t, then this interval cannot splinter into two or more non-contiguous intervals of activated responses at $t + 1$. Together, these results imply both contiguous activation and sequential transitions.