If Hot Coherence is Rational, then How? Review of Paul Thagard (2006) "Hot Thought: Mechanisms and Applications of Emotional Coherence"

Iris van Rooij
Eindhoven University of Technology
Eindhoven, The Netherlands

Can human beliefs and inferences be understood as a form of coherence maximization? This question underlies much of the research that computational philosopher Paul Thagard has performed over the past two decades. In his most recent book, *Hot Thought*, Thagard continues his investigation of the explanatory value of the coherence theory by bringing in the idea that human thought may be "hot" in the sense that it is emotionally colored through and through. The book is a natural extension of Thagard's earlier book *Coherence in Thought and Action* and the theory of emotional coherence, called HOTCO, discussed therein. In *Hot Thought*, the emotional coherence theory is thoroughly revised in the form of HOTCO2, in which emotions are attributed causal thinking powers, while HOTCO modeled emotions as mere epiphenomena of rational or otherwise "cold" thinking processes. Thagard's coherence theory can be informally characterized as follows: Mental representations can cohere (fit together) or incohere (resist fitting together). If two representational elements (say, propositions) *p* and *q* cohere (incohere) then believing *p* will tend to increase (decrease) believe in *q*, and vice versa. The *emotional* variant of coherence theory furthermore introduces valence values associated with *p* and *q*. Valences and beliefs are assumed to interact in complex ways via coherence and incoherence relations so as to bring about a stable pattern of beliefs and disbeliefs.

In *Hot Thought*, Thagard does an admirable job building the case for applicability and generality of his theory. He shows, among other things, how the legal notion of "reasonable doubt", group consensus, jury decision-making, the "will to believe in God", self-deception, and scientists decisions about which research goals to pursue and which scientific hypotheses to believe, all can be conceptualized as outcomes of an unconscious emotional (in)coherence computation. As such, the emotional coherence theory presents an important alternative and competitor to Expected Utility Theory and Bayesian accounts of human thinking—approaches that Thagard explicitly rejects as being psychologically unrealistic. Thagard claims psychological realism for his own theory, not only as a *descriptive* model of how people think, but also as a *normative* model of how people should think. According to Thagard, when thinkers are careful and take into account all relevant information, then the maximization of emotional coherence will lead them to have "rational" beliefs in the sense that the beliefs are justified and conducive of truth.

The story that Thagard has to tell is without a doubt a "good story", but whether or not goodness of story is indicative of truth, or even of truthlikeness, remains a controversial issue (Dawes, 2001). Moreover, whether or not the reader will be convinced by this story depends, I suspect, in large part on the intuitions that the reader brings to the reading of the book. My own intuitions are very much in line with Thagard's. I am sympathetic to the idea that much of human thinking can be modeled as a form of coherence maximization. I even find it plausible that reasoned or rational judgments can be "hot" at times. In fact, I suspect that Thagard may be very much on the right track and reading *Hot Thought* strengthens this conviction. Yet I cannot help having the feeling that this is merely because Thagard's pre-theoretical story resonates well with me. This need not be a problem for Thagard. After all, in his opinion a person should simply

judge to what extent the content of the book is emotionally coherent to decide whether or not to accept it, and apparently it is of high emotional coherence against the background of my system of beliefs. I should thus rationally infer that what Thagard is telling me is true, close to the truth, or otherwise acceptable. I may have been tempted to accept this argument if it weren't for the little devils that Thagard unleashes with the details of his formal model of emotional coherence and its questionable normative status.

To explain my concerns, I recapitulate the formalisms underlying Thagard's emotional coherence theory here. To start, HOTCO2 models representational elements as nodes in a belief network, with connections modeling (in)coherence relations between pairs of elements. Each element p_i is assumed to have two associated values, an activation value $-1 \leq a_i \leq 1$ and a valence value $-1 \leq v_i \leq 1$, representing *degree of belief* and *emotional attitude* respectively. More precisely,

1. if a_i is positive then this means that the proposition expressed by p_i is believed to degree a_i, and if a_i is negative then this means that p_i is disbelieved to degree $-a_i$;
2. if v_i is positive then this means that the person has a positive emotional attitude towards p_i of strength v_i, if v_i is negative then this means that the person has a negative emotional attitude towards p_i of strength $-v_i$.

For each connection (p_i, p_j) there is a weight $-1 \leq w_{ij} \leq 1$, representing the degree to which the elements *cohere*:

3. if w_{ij} is positive then this means that p_i and p_j cohere with strength w_{ij}, if w_{ij} is negative then this means that p_i and p_j incohere with strength $-w_{ij}$.

Lastly, HOTCO2 assumes that emotional coherence maximization consists in an updating of activation and valence values according to the following updating rules:

$$a_j(t+1) = a_j(t) + f\left(\sum_i w_{ij} a_i(t) + \sum_i w_{ij} v_i(t) a_i(t)\right) \qquad (1)$$

and

$$v_j(t+1) = v_j(t) + f\left(\sum_i w_{ij} v_i(t) a_i(t)\right) \qquad (2),$$

where $f(.)$ is a non-decreasing function that scales the inputs to a_j and v_j such that these activations and valences asymptote at a value between –1 and 1. The updating process continues until the system settles in a stable pattern. At the end of the process, emotional coherence—as Thagard defines it—has been maximized and the activation values represent justified or rational degree of belief.

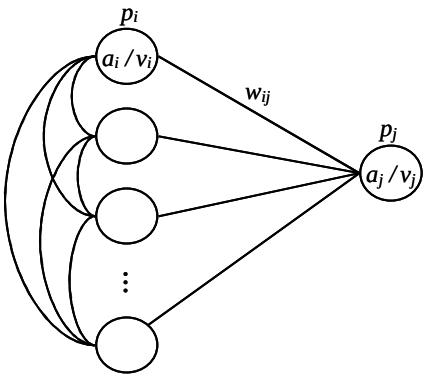

Figure 1. Illustration of a belief network: Two nodes in the network are labeled p_i and p_j, having activation values a_i and a_j and valence values v_i and v_j respectively. The link connecting any two nodes p_i and p_j has weight w_{ij}. Node p_j receives activation and valence input from p_i and all other nodes connected to it (see also Table 1).

A question arises at this point: Why does Thagard propose rules (1) and (2)? In other words, why does he think it is most coherent to update one's beliefs according to these rules? Nowhere in the book could I find a justification; not even an argument for why these updating rules are reasonable or intuitive. Thagard may find it obvious, or he may find the issue unimportant, as he hides the formal discussion of these technical details in an Appendix to Chapter 3 of his book. He may also find it more useful to focus in the main text of his book on the informal intuitions underlying his model rather than the formal details. But there is a problem with this approach: the simulation results reported in the main text of the book, and therefore the arguments for applicability of the model to different domains, depend crucially on the technical details of the model. If HOTCO2 fails to capture the intuitive notion of emotional coherence then the reported simulation results are irrelevant for assessing the applicability of that notion to different domains. After all, in that case there is no principled relationship between the HOTCO2 model and the pre-theoretical notion of 'emotional coherence'.

In this review, I investigate if we can possibly discern a rationale for the updating rules adopted in the HOTCO2 model. In particular, I consider two possible ways in which these rules could be justified. First, I consider the possibility that the rules implement intuitions about how degree of belief and valence interact locally in a belief network (called *local justification*). Second, I consider the possibility that belief updating according to the proposed rules leads to a global pattern that intuitively implements maximum emotional coherence (called *global justification*).

Consider the belief network in Figure 1. To investigate the local justifiability of the model's updating rules we focus on the input that p_j receives from p_i. Table 1 presents an overview of the activation and valence inputs to p_j from p_i for different activation/valence values of p_i. For simplicity, we consider the values –1 and +1 only. Furthermore, we distinguish three types of inputs from p_i to p_j; the valence input $w_{ij}v_ia_i$ and the "hot" activation input $w_{ij}a_i + w_{ij}v_ia_i$ assumed by the HOTCO2 model, and the "cold" activation input $w_{ij}a_i$ assumed by Thagard's original 'cold" coherence model. Table 1 allows us to assess if HOTCO2's updating rules implement our intuitions about how belief, coherence and emotion (should) *locally* interact.

Table 1. Cold, hot and valence input from p_i to p_j as a function of w_{ij}, v_i, and a_i

			input from p_i to p_j		
			cold activation	valence	hot activation
w_{ij}	v_i	a_i	$w_{ij}a_i$	$w_{ij}v_ia_i$	$w_{ij}a_i + w_{ij}v_ia_i$
1	1	1	1	1	2
1	1	-1	-1	-1	-2
1	-1	1	1	-1	0
-1	1	1	-1	-1	-2
1	-1	-1	-1	1	0
-1	1	-1	1	1	2
-1	-1	1	-1	1	0
-1	-1	-1	1	-1	0

I have always found it a counterintuitive aspect of the cold coherence model of Thagard (2000) that disbelieving a proposition p_i enhances the belief in other propositions with which it incoheres[1] (see a_iw_{ij} in rows 6 and 8 in Table 1), and as it turns out, this counterintuitive property is inherited by the hot coherence model (at least, for $v_i > a_i < 0$). Also, the model assumes that valence, belief, and coherence combine multiplicatively to produce valence inputs to other propositions. This captures the intuition (*a la* cognitive dissonance theory) that a positively valued belief tends to induce negative attitudes in beliefs it incoheres with (see $w_{ij}v_ia_i$ in row 4 of Table 1), but at the same time it has some counterintuitive consequences; e.g, that disbelieving a proposition p_i with negative valence tends to induce positive attitudes towards propositions with which p_i coheres (see $w_{ij}v_ia_i$ in row 5 of Table 1), as well as that disbelieving a proposition p_i with positive valence tends to induce positive attitudes towards propositions with which p_i incoheres (see $w_{ij}v_ia_i$ in row 6 of Table 1). Furthermore, the hot coherence model assumes that cold belief and valence input combine additively to produce hot belief inputs (compare columns 4 and 5 with column 6 in Table 1). Even if these proposed interactions between belief, emotion and coherence are truly descriptive of how humans update their beliefs and attitudes, it seems hard to imagine how they can be given a normative justification.

Though rationality may not be apparent at the local level, the *global* outcome of the complex and non-linear local interactions in a belief network may still have normative status. Perhaps a justification for updating rules (1) and (2) should be sought in the global activation patterns in which HOTCO2 tends to settle. But what special normative properties do the activation patterns produced by HOTCO2 have? Remarks by Thagard on this point are rather informal and come down to the claim that HOTCO2 implements a method for *maximizing the satisfaction of multiple cognitive and emotional constraints* (e.g., pp. 20, 30, and 161). Unfortunately, Thagard does not specify what constraint satisfaction means in the HOTCO2 model. I will try to reconstruct what I think he has in mind, by analogy to constraint satisfaction as it was defined in his cold coherence model (Thagard, 2000). In the cold coherence model, which assumes the following updating rule:

[1] It seems to imply that one can artificially and arbitrarily bump up one's belief in p by introducing blatantly false propositions q_1, q_2, …., q_n that all incohere with p.

$$a_j(t+1) = a_j(t) + f\left(\sum_i w_{ij} a_i(t)\right), \quad (3)$$

maximum constraint satisfaction is defined as computing an activation pattern that maximizes *(cold) harmony*:

$$H_{cold} = \sum_i \sum_j w_{ij} a_i a_j \quad (4)$$

By analogy, constraint satisfaction in the *hot* coherence model may be taken to mean maximizing *hot harmony*, defined as follows:

$$H_{hot} = \sum_i \left(\sum_j w_{ij} a_i a_j + \sum_j w_{ij} v_i a_i a_j + \sum_j w_{ij} v_i a_i v_j \right) \quad (5)$$

There seem to be at least two problems with this idea however. First of all, if maximizing H_{hot} is indeed what maximizing emotional coherence amounts to, then the HOTCO2 model does not explain how this maximization is done by human minds. After all, updating rules (1) and (2) do not ensure that H_{hot} is maximized (in the same way that updating rule (3) does not ensure that H_{cold} is maximized; see also Milgram, 2000; van Rooij & Wright, 2006). This failure of the HOTCO2 model is already apparent for small networks. Consider, for example, a belief network with only two nodes p_i and p_j connected by a negative weight $w_{ij} = -1$, and let valences v_i and v_j both be preset at +1 (i.e., there is a positive attitude towards the two beliefs, but the beliefs incohere). For this network H_{hot} is maximized if $a_i = -1$ and $a_j = +1$, in which case $H_{\text{hot}} = 2(w_{ij}a_ia_j) + w_{ij}v_ia_ia_j + w_{ij}v_ja_ia_j + w_{ij}v_ia_iv_j + w_{ij}v_ia_jv_j = 2 + 1 + 1 + 1 - 1 = 4$. If we were to update activations according to rules (1) and (2), using the parameter settings adopted by Thagard, we would end up with (small) positive values for both a_i and a_j, which leads to a *negative* value for H_{hot}.

Even if this first problem could somehow be overcome by resetting the model's parameters, a second problem remains: If Equation (5) correctly describes hot harmony, then maximizing hot harmony is equivalent to maximizing cold harmony, at least as far as belief fixation is concerned. The reason is that a constraint w_{ij} in a belief network contributes maximally to H_{hot} if only if one of the following conditions is met:

- $w_{ij} = +1$ and $a_i = a_j = 1$;
- $w_{ij} = -1$, $a_i = 1$ and $a_j = -1$; or
- $w_{ij} = -1$, $a_i = -1$ and $a_j = 1$,

which are exactly the same conditions under which H_{cold} is maximized (Thagard & Verbeurgt, 1998).[2] The equivalence is admittedly not a problem in and of itself, but it makes it problematic to distinguish between cold and hot forms of rationality, both empirically and normatively. Furthermore, since the normative status of belief fixation by maximizing H_{cold} is far from established, and even questionable (Milgram, 2000; van Rooij & Wright, 2006), there is so far no basis for believing that maximization of H_{hot} is a normative model of belief fixation.

[2] I discovered this equivalence when exhaustively trying out all possible values for a_i and a_j in a two-node network that maximize H_{hot} for different input values for a_i, a_j, v_i and v_j, but the result can probably also be derived analytically.

Although my review may seem rather technical, it is certainly not my intention to underplay the main points and contributions made by *Hot Thought*. Rather, what I hope to achieve is an awareness of the technical problems and subtleties of computational modeling of "hot cognition", and in particular "hot rationality". What I found lacking in *Hot Thought* was a rigorous validation and analysis of the proposed model and a self-critical stance towards the simulation results, their robustness and interpretation. As a result some counterintuitive and non-normative aspects of the model have gone unnoticed. Be that as it may, *Hot Thought* does convincingly argue for the need to investigate alternatives to classical decision theories of human inference based on the notion of emotional coherence. I applaud Thagard's efforts in this direction and believe investigations in the same direction should grow in number and diversity (along the lines of Chapters 3, 5 and 6 in *Hot Thought*, or otherwise). I do hope to have illustrated the need for rigorous computational analyses in this pursuit. The mere availability of a computational model that fits human belief data is insufficient if one's goal is to explain how human belief fixation is both computationally feasible and justified. To achieve the latter the computational model should satisfy the right kind of explanatory constraints. Thagard's emotional coherence model does not (yet) seem to satisfy those constraints.

References

Dawes, R. (2001). *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally*.

Millgram, E. (2000). Coherence: The price of the ticket. *Journal of Philosophy*, 97, 82-93.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science, 22*, 1–24.

van Rooij, I. & Wright, C. (2006). The incoherence of heuristically explaining coherence. In R. Sun & N. Miyake (Eds.), *Proceedings of 28th Annual Conference of the Cognitive Science Society* (p. 2622).

Figures

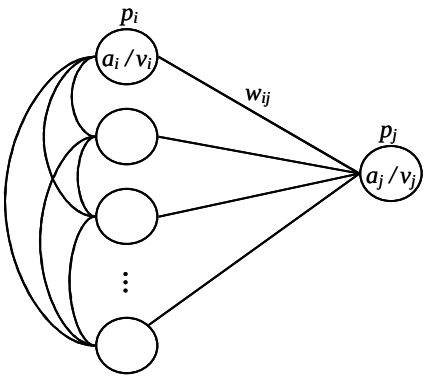

Figure 1. Illustration of a belief network: Two nodes in the network are labeled p_i and p_j, having activation values a_i and a_j and valence values v_i and v_j respectively. The link connecting any two nodes p_i and p_j has weight w_{ij}. Node p_j receives activation and valence input from p_i and all other nodes connected to it (see also Table 1).

Tables

Table 1. Cold, hot and valence input from p_i to p_j as a function of w_{ij}, v_i, and a_i

			input from p_i to p_j		
			cold activation	valence	hot activation
w_{ij}	v_i	a_i	$w_{ij}a_i$	$w_{ij}v_ia_i$	$w_{ij}a_i + w_{ij}v_ia_i$
1	1	1	1	1	2
1	1	-1	-1	-1	-2
1	-1	1	1	-1	0
-1	1	1	-1	-1	-2
1	-1	-1	-1	1	0
-1	1	-1	1	1	2
-1	-1	1	-1	1	0
-1	-1	-1	1	-1	0