May 14th, 9:00 AM - May 17th, 5:00 PM

Connectionist Coherence and Moral Reasoning

Marcello Guarini
University of Windsor

Title: Connectionist Coherence and Moral Reasoning
Author: Marcello Guarini
Commentary: A. Bailey

---

1. Moral Reasoning as Constraint Satisfaction

To understand Thagard's views on the nature of moral reasoning, we must first examine what he means by "coherence." The basic idea is that when we try to make sense of *things*, we select an interpretation that *fits* the available information better than any available alternative. The *things* we are trying to make sense of and the way we explain what *fits the available information* depends on the domain being modelled. Thagard has developed a general account of coherence that applies to a variety of domains. The following principles summarize, in a very general way, Thagard's views on setting up a coherence problem.

> 1. Elements are representations such as concepts, propositions, parts of images, goals, actions, and so on.
> 2. Elements can cohere (fit together) or incohere (resist fitting together). Coherence relations include explanation, deduction, facilitation, association, and so on. Incoherence relations include inconsistency, incompatibility, and negative association.
> 3. If two elements cohere, there is a positive constraint between them. If two elements incohere, there is a negative constraint between them.
> 4. Elements are to be divided into ones that are accepted and ones that are rejected.
> 5. A positive constraint between two elements can be satisfied either by accepting both of the elements or by rejecting both of the elements.
> 6. A negative constraint between two elements can be satisfied only by accepting one element and rejecting the other.
> 7. The coherence problem consists of dividing a set of elements into accepted and rejected sets in a way that satisfies the most constraints (Thagard and Verbeurgt 1998, 2-3).

There are two senses of "coherence" at work here. First, two elements of a domain may be said to cohere or incohere with one another – this is local coherence. Second, when an algorithm settles on a set of many elements by satisfying as many constraints as possible between cohering

and incohering pairs of elements, this is global coherence. In a variety of domains, Thagard conceives inferences to best explanation as an attempt to maximize global coherence.

Thagard has developed a theory of explanatory coherence that he uses to account for theory selection in science. There have been many instances in the history of science where theories have been in competition with one another, and one theory was selected over another. Thagard views theory selection in science as a coherence problem. Roughly, the *elements* in this type of coherence problem consist of (a) the propositions making up the competing theories and (b) the propositions expressing the evidential support for the theory. Propositional elements are understood to correspond to the units of a neural network; positive constraints correspond to excitatory links: negative constraints to inhibitory links; an accepted element corresponds to a positively activated unit, and a rejected element corresponds to a negatively activated unit. ECHO is a program that has been used to model a variety of theory conflicts from the history of science – Ptolemaic vs. Copernican views of astronomy, phlogiston versus oxygen chemistry, and so on. In every case, using the same parameter values, ECHO settles on the correct theory. Thagard conceives of theory selection in science as inference to the best explanation, and inference to the best explanation is understood as maximizing global coherence. A brief example is in order.

Say we have two hypotheses that compete with one another, H1 and H2. Let us say that E1, E2, E3, E4, and E5 are propositions expressing evidence. Consider figure 1. A thick line expresses an inhibitory link; a thin line expresses an excitatory link. Since the two hypotheses are in competition with one another, there is an inhibitory link between them. Excitatory links are established between a hypothesis and the evidence it explains. Also, there is a special evidence unit (SEU) to which all pieces of evidence have an excitatory link; this establishes a kind of priority for the data. The neural net resulting from putting a neuron at each node is a settling net (as opposed to, say, a feed forward net). Every neuron or unit can be thought of as firing simultaneously. The activation value of each unit j, represented by a_j — that can range between -1 (rejected) and 1 (accepted) — is calculated as a function of the old value, a_i, of every unit i linked to j. The following equation is used to carry out the calculations.

$$a_j(t+1) = a_j(t)(1-\theta) + \begin{cases} net_j(max-a_j(t)) & \text{if } net_j>0 \\ net_j(a_j(a_j(t)-min) & otherwise \end{cases}$$

The variable θ is a *decay parameter* that decrements each unit at every cycle; *min* is the minimum activation (-1); *max* is the maximum activation (1), and net_j is the net input to a unit, defined by the following equation:

$$net_j = \sum_i w_{ij} a_i(t)$$

This process of updating unit activation is iterated until the network settles (if it settles). When the network settles, some units will have positive activation (representing accepted propositions), and some will have negative activation (representing rejected propositions). While the SEU unit is clamped (so that it is always active or accepted), it is possible in some networks (depending on how the excitatory and inhibitory links are set up) that not all the data or evidence units will remain active when (and if) the network settles. While the SEU gives data priority, it does not guarantee they will all be accepted.

Figure one grossly oversimplifies Thagard's account of global explanatory coherence since it does not consider the role that is granted to simplicity, unification, and other factors. However, it is enough to give us a feel for what Thagard has in mind when he writes about global coherence. Roughly, the collection of elements (which are statements of evidence and hypotheses in the example at hand) that satisfies the most constraints in a network of elements has the most overall or global coherence.

One more point is worth making about Thagard's methodology in studying scientific reasoning: it is both descriptive and prescriptive:

> I have coined a new term to describe an approach that is intended to be both descriptive and prescriptive (normative). I shall say that a model is "biscriptive" if it describes how people make inferences when they are in accord with the best practices compatible with their cognitive capacities. Unlike a purely prescriptive approach, a biscriptive approach is intimately related to human performance. But unlike a purely descriptive approach, biscriptive models can be used to criticize and improve human performance (Thagard1992, 97).

Both the coherence approach to modelling reasoning and the biscriptive methodology are applied to Thagard's attempt to model moral reasoning.

In "Ethical Coherence," Thagard argues that the coherence at issue in ethical reasoning is a kind of meta-coherence, or coherence of coherences. He identifies four types of coherence that figure into the overall coherence of ethical reasoning: explanatory, deductive, deliberative, and analogical. Above, we saw how his account of explanatory coherence can be used to model scientific reasoning. It is also one of the four parts of moral reasoning. The reason for including this type of coherence in the account of moral reasoning is that some normative principles are tied to empirical claims. For example, the general principle that capital punishment is acceptable may be argued to depend on the deterrent effect that it has. But whether capital punishment has a deterrent effect is a largely empirical question. ECHO can be set up to take general principles, particular judgements, and empirical evidence as elements in an attempt to maximize explanatory coherence. Figure 2 provides an example of a mutually constraining package of propositions. The principle that practices preventing serious crimes are good together with the claim that capital punishment is a practice that prevents serious crime entails that capital punishment is good. Deductive coherence is about finding a reasonable fit between general principles and particular judgements. ECHO can be set up so that it takes principles and particular judgements as elements. For example, if a general principle entails a particular judgement, then there is a positive constraint between the principle and the judgement. Inconsistencies yield negative constraints. ECHO can be run on a body of principles and judgements to maximize coherence, yielding as good a fit as possible between principles and judgements.

Deliberative coherence is about finding a reasonable fit between judgements and goals. Paul Thagard and Elijah Millgram have developed a coherence theory of decision making that takes as its elements actions and goals (Thagard and Millgram 1995; and Millgram and Thagard 1996). The primary positive constraint between these elements is facilitation, and the negative constraint is the incompatibility of an action with a goal. For example, the goal of saving tax

dollars may be facilitated by capital punishment, whereas imprisoning individuals for decades may not facilitate that goal. There is more:

> Just as explanatory coherence gives some priority to propositions that state empirical evidence, so deliberative coherence gives some priority to intrinsic goals, ones that an agent has for basic biological or social reasons rather than because they facilitate other higher goals. But just as empirical evidence can be overridden for reasons of explanatory coherence, intrinsic goals can also be revised and overridden for reasons of deliberative coherence, which evaluates intrinsic goals (final ends) as well as instrumental goals and actions (Thagard 1998, 411).

DECO is a program created by Thagard and Millgram that carries out the computation of coherence given goals and actions as elements. Intrinsic goals are given a kind of defeasible priority in manner not unlike the way data are given priority in explanatory coherence.

Analogical coherence is about finding a reasonable fit between judgements of some cases with the judgements of other cases. Interlocutors sometimes appeal to an agreed upon case (the source) to argue that some disputed case (the target) should be treated in the same way. For example, Judith Thomson (1971) has argued that some cases of forcing a woman to go through with a pregnancy are similar to a hypothetical case where a person is kidnapped and forced to stay hooked up to an individual to keep that individual alive. She suggests that if the latter is unacceptable, then so are the former. Keith Holyoak and Paul Thagard (1995) have developed a coherence approach to determining the strength (or lack thereof) of analogical correspondence between two cases. The program that implements their approach to analogical mapping is called ACME, and it takes as its elements hypotheses about what features of the source and target cases correspond to one another. In their multiconstraint account of analogy, positive constraints are based on semantic similarity, visual similarity, syntactic similarity, and the purpose of the proposed analogue. A connectionist algorithm is then used to try to maximize the satisfaction of these constraints (which is to say, to maximize analogical coherence).

Thagard applies his multicoherence account of moral reasoning to the considerations that some may have in contemplating the appropriate punishment for Paul Bernardo, a Canadian who was convicted of the prolonged sexual torture and murder of two young women. Canadian law does not allow for capital punishment, so he was sentenced to life in prison. Some who have long argued against capital punishment believed that Bernardo deserved to be executed, a view which did not cohere with all their other views. Figure 2 captures some of the deductive, explanatory, deliberative, and analogical considerations that are at work in the consideration of Bernardo's fate and capital punishment.[1] Since ECHO, DECO, and ACME use the same constraint satisfaction algorithm for maximizing coherence, this algorithm can be applied to the constraint network in figure two to determine which elements (propositions) should remain activated and which should not (Thagard 1998, 415).

2. Some Problems

While I acknowledge that Thagard's multiconstraint account of coherence in moral reasoning is far more sophisticated than most of the vague accounts of reflective equilibrium which have been offered thus far, there are places where I think he may have overstated its strengths. He claims that his account overcomes

> the two major problems of foundationalist approaches to ethics and epistemology. The first problem is that, for epistemology as for ethics, no one has ever been able to find a set of foundations that even come close to receiving general assent. . . . The second problem is that proposed foundations are rarely substantial enough to support an attractive epistemic or ethical edifice, so that foundationalism degenerates into skepticism (Thagard 1998, 418-419).

In the rest of this section, I will argue that these conclusions are, *at best*, premature.

As we noted earlier on, Thagard takes a biscriptive approach to modelling scientific reasoning. Not surprisingly, he takes the same approach to modelling moral reasoning — the theory of ethical coherence describes how people engage in moral thinking when they do it at their best (Thagard 1998, 417). One problem with this is that unlike the sciences, where there is a quite a bit agreement regarding which theories should have won when theories competed for acceptance, it is not clear that the same level of agreement exists regarding conflicting ethical views. Without the required agreement, neither of the two traditional problems identified by Thagard for foundationalism in ethics is overcome.

There are a variety of parameters at work in the consideration of each type of coherence. For example, the weights attached to excitatory and inhibitory links may be set in different ways. Even if there is agreement on how a particular set of disagreements is to be resolved, that agreement will underdetermine the parameter selection for each type of coherence. As a result, there may not be agreement on how to handle new disagreements since the different parameter values compatible with the agreed upon resolution of past disagreements may not yield identical solutions for future disagreements. This is one of the underdetermination problems that afflict the multi-constraint account of ethical coherence. There is a second underdetermination problem.

There is a real issue of how the different types of coherences should be weighted. It is far from obvious that each type of coherence should count equally. As Thagard himself points out, a devout Kantian may put no weight on empirical considerations, and a utilitarian may put no weight on non-empirical considerations (Thagard 1998, 418). R.M. Hare, no doubt, would put little or no weight on analogical considerations since they frequently appeal to substantive intuitions that are simply assumed to be correct by the participants of a dialogue, and Hare (1981, 11-15) rejects the use of such assumptions. In discussing these types of concerns, Thagard makes the following remarks:

> My response is first to point out that the extreme versions of both these approaches [Kantian and utilitarian] have familiar incoherences with most people's ethical judgements, and second to

> point to the multifarious nature of actual ethical arguments that embrace different kinds of ethical concerns, including both Kantian and utilitarian ones. I do not have an algorithm for establishing the weights on people's constraints, only the hope that, once discussion establishes a common set of constraints, coherence algorithms can yield consensus (Thagard 1998, 418).

Not bad — but not enough to claim that his theory of ethical coherence overcomes the two traditional problems of foundationalism in ethics. Assuming the legitimacy of Thagard's points, the most that follows is that *some* weight must be given to both Kantian and utilitarian types of considerations. What is not clear is how much weight must be assigned to each of the four types of coherences in calculating overall coherence. Moreover, even if discussion should yield agreement on which types of coherence should be involved in moral reasoning, and even if that discussion should yield agreement on a common set of correct answers in past moral disagreements, all that agreement may still underdetermine the weights to be assigned to the different types of coherence in the computation of overall coherence. In other words, different assignments of weights for the different types of coherence may yield the same answers for an agreed upon body of successfully resolved disputes but not yield the same answers when new disputes arise. This is the second underdetermination problem referred to above.

The two underdetermination problems and the concerns over finding significant agreement over what counts as good moral reasoning make it far from obvious that the two traditional problems in foundationalist ethics have been overcome. There is another problem as well. Consider the constraint network presented in figure three; it lays out some of the considerations which are relevant to the debate over capital punishment in the mind of an individual who has the intuition that Paul Bernardo should be executed. Propositions connected by thin lines are positively constrained (or tend to be accepted together) and propositions connected by thick lines are negatively constrained (or tend to be rejected together). There is a real issue here of what should count as one proposition and what should count as many. This problem can be seen in a number of ways, but I will focus on how it applies to deductive coherence. One normative deduction can be broken down into many, and this affects the results of reasoning in Thagard's model. Consider the following:

(1) An action is right only if it maximizes happiness;
(2) Jack broke a promise when doing so did not maximize happiness;
(3) Jack did not do the right thing;
(4) breaking a promise is right only if it maximizes happiness, and
(5) keeping a promise is right only if it maximizes happiness.

Clearly, (1) and (2) jointly entail (3). Just as clearly, (4) and (2) jointly entail (3); and (5) and (2) jointly entail (3). According to Thagard's theory of coherence, if other things are equal, a constraint network having (4) and (2) positively connected to (3) as well as having (5) and (2) positively connected to (3) will have greater coherence (or acceptability) than a network that has only (1) and (2) connected to (3). Yet this seems like the wrong answer. Even if we start with a moral theory that has only one moral principle, we can generate other principles which follow from it to get the result that more entailments for a proposition leads to greater acceptability or coherence (other things being equal). It is worth keeping in mind that for Thagard, *acceptability*

is cashed in terms of *coherence*. Entailment is the strongest type of logical support, so if (1) and (2) entail (3), it is not clear how support for (3) could be stronger. If someone should be tempted by the idea that more entailments yield greater support (or coherence), consider the following.

(6) This shirt contains the colour red.
(7) This shirt is coloured.
(8) This shirt contains the colour blue.

Clearly, (6) and (8) each entail (7), but is the argument from the conjunction of (6) and (8) to (7) any stronger than the argument from (6) to (7) or the argument from (8) to (7)?
No. One of the consequences of Thagard's coherence theory of moral reasoning is that (other things being equal) more positive linkage is better, even when support is already maximal.

The obvious way to take issue with the above argument is to question the exceptionless nature of (1), (4), and (5). It might be said that most, if not all, these propositions contain implicit ceteris paribus clauses. As I see it, either (a) implicit ceteris paribus clauses are eliminable from some moral principles, or (b) they are not. If (a) is the case, then we can rerun the above argument using claims that have no implicit ceteris paribus clauses. If (b) is the case, then there is no such thing as deductive coherence in moral reasoning since all principles have ineliminable ceteris paribus clauses, which precludes deduction, which means that the deductive constraint would have to be eliminated from Thagard's multi-constraint theory. Whether we take path (a) or (b), there is a problem.

The final and most important problem to be raised in this paper concerns our inability to argumentatively access global coherence. Presumably, if belief set B2 has a greater level of global coherence than belief set B1, then we are justified in believing B2 over B1. The problem with this is that we do not – as a general rule – have argumentative access to global coherence. In order to argue for B2 over B1, there has to be reason for accepting the premise "B2 has greater overall coherence than B1." However, our (conscious) memory is far too limited to actually try to work out, for any significant number of beliefs, its overall coherence rating. The average person cannot use the equations mentioned earlier in this paper together with other equations to work out which set of beliefs has a higher coherence rating. This has implications for Thagard's biscriptive methodology. If we are trying to capture how we reason when we reason at our best, and if when we reason at our best we cannot make an argument for B2 over B1 on the grounds that the required coherence ratings are not available to us, then it is difficult to believe that the connectionist coherence account of moral reasoning presented herein is plausible. Notice: it will do no good to postulate that a suitably powerful computer could be used to help us carry out the calculations since part of what Thagard is trying to capture is what we have historically regarded as good reasoning by good reasoners, and such individuals have not had access to powerful computing devices to engage in what Thagard agrees is good reasoning. The inability to derive coherence ratings (without the assistance of a computational device or a pencil, plenty of paper, and plenty of time) for alternate sets of beliefs leads to the inability to *argue* for one set as superior to another. Moreover, the problem cannot be solved by simply asserting that it is enough that some neural net in the brain settles on the belief set that has greatest coherence without our having any argumentative access to the coherence rating of the belief set. Without the ability to *argue* that one set of beliefs is superior to another, it is difficult to see how Thagard's theory could muster sufficient prescriptive force to overcome the traditional problems of foundationalist approaches to ethics, and that was one of his stated goals.

I take the above arguments to be a friendly critique of Thagard's position. No doubt, some readers may think that with friends like Guarini, Thagard is hardly in need of enemies. However, it should be noted that none of the arguments presented herein call into question the very possibility of shedding light on the nature of moral reasoning by making use of computational models of reasoning. The point has simply been that the model of reasoning presented in "Ethical Coherence" suffers from non-trivial problems. Perhaps these problems can be overcome in a slightly modified model, or perhaps significantly different computational models will prove more fruitful. While the latter of these options strikes me as more plausible, either of them is compatible with the view that attempting to computationally model moral reasoning might help us to better understand it (both what it is, and what it is not). On that point, I *am* in agreement with Thagard.

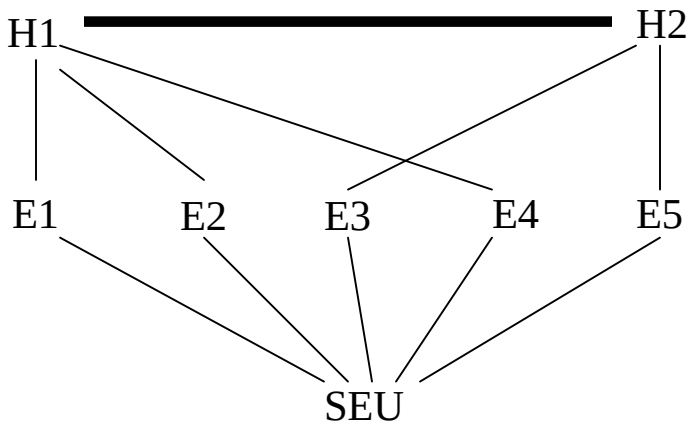

Figure 1

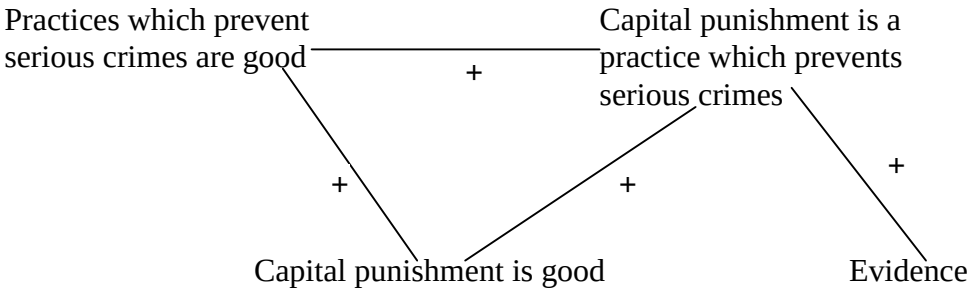

Figure 2

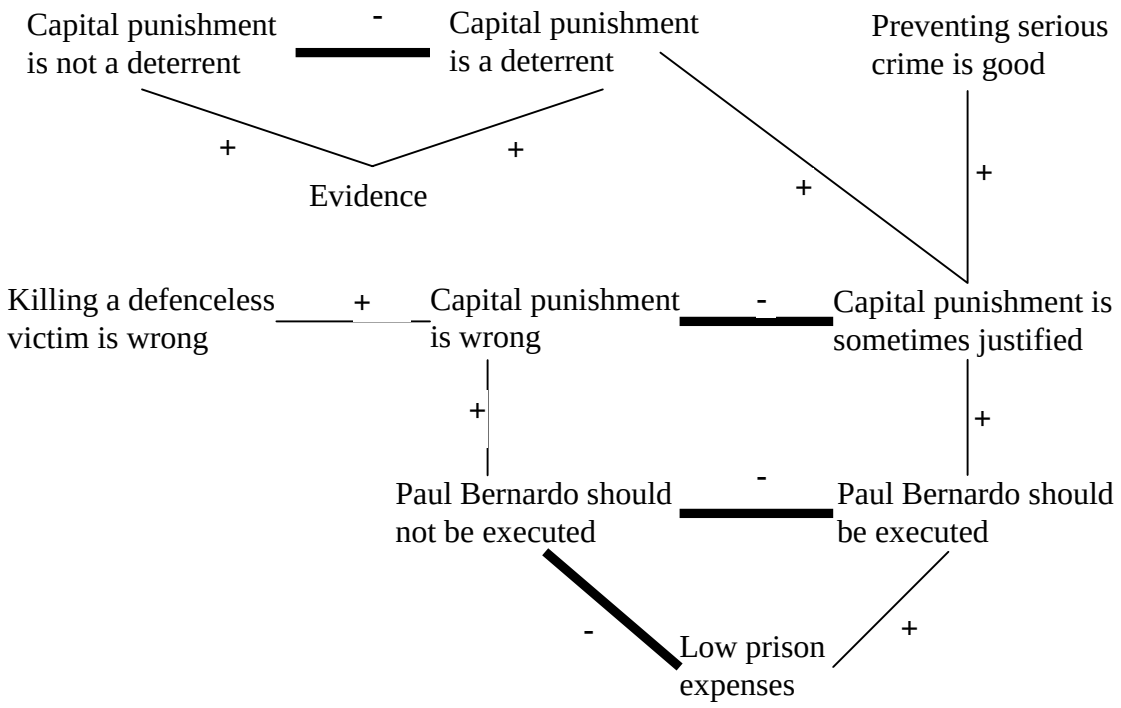

Figure 3

Notes

[1] Thagard frequently abbreviates in presenting constraint networks. For example, strictly speaking, "Paul Bernardo should not be executed" cannot be deduced from "Capital punishment is wrong," which are two of the propositions in his constraint network. The latter does entail the former, but the proposition "Executing Paul Bernardo is an instance of capital punishment" needs to be added in order for the deduction to go through. The last proposition is not in the constraint network; I take this to be an abbreviation. (Similarly, in the argument, "This is red; therefore, this is coloured," the first claim entails the second, but "red objects are coloured" has to be added if the argument is to be understood deductively.) Since entailment can often be seen without a complete deductive reconstruction, and since complete deductive reconstruction is often not required and can make the presentation of a point tedious, it is reasonable to abbreviate deductions with entailments that leave out some propositions that are required for the deductions to hold. I will follow this strategy myself, often speaking of entailments and assuming that it is understood that propositions need to be added for the deductions to hold. However, it should be kept in mind that Thagard is claiming to model *deductive* coherence.

References

Eliasmith, C. and P. Thagard. 1997. "Waves, Particles, and Explanatory Coherence," *British Journal for the Philosophy of Science* 48**:** 1-19.

Hare, R. M. 1981. *Moral Thinking: Its Levels, Methods, and Point*. Oxford: Clarendon Press.

Holyoak, K.J. and P. Thagard. 1995. *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press/Bradford Books.

McClelland, J.L., and D.E. Rumelhart. 1981. "An Interactive Activation Model of Context Effects in Letter Perception: Part 1: An Account of Basic Findings," *Psychological Review* 88: 375-407.

Millgram, E. and P. Thagard. 1996. "Deliberative Coherence," *Synthese* 108: 63-88.

Thagard, P. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Thagard, P. 1998. "Ethical Coherence," *Philosophical Psychology* 11: 405-22.

Thagard, P. and E. Millgram. 1995. "Inference to the Best Plan: A Coherence Theory of Decision," in *Goal-driven Learning*. Cambridge, MA: MIT Press, pp. 439-54.

Thagard, P. and K. Verbeurgt. 1998. "Coherence as Constraint Satisfaction," *Cognitive Science* 22: 1-24.

Thomson, J.J. 1971. "A Defense of Abortion," *Philosophy and Public Affairs* 1: 47-66.