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

Commentary on Guarini

Andrew Bailey

Author: Andrew Bailey
Commentary on: M. Guarini's "Connectionist Coherence and Moral Reasoning"

Marcello Guarini's paper is a critical examination of Paul Thagard's computational model of moral reasoning. As Guarini explains, Thagard's basic idea is that we can, in general, model many of our rational deliberations as a process of settling on an *interpretation* of a particular domain that *fits* the available information better than any available alternative. The important notion of 'fitting' is cashed out by Thagard in terms of *coherence*. The "elements" of a domain—such as a particular theory-selection problem in science, an ethical dilemma, or a difficult choice between competing job offers—are treated by Thagard as a set of representations. These representations stand in various relations of *local* coherence, according to Thagard: some of them cohere with other elements, in virtue of such more primitive relations as "explanation, deduction, facilitation, association, and so on" (Thagard and Verbeurgt 1998, 3), and others *in*cohere, because of their relations of "inconsistency, incompatibility, and negative association" (Thagard and Verbeurgt 1998, 3). This set of local coherences between the elements of a domain can be modelled as giving rise to a set of *constraints*: where two elements cohere, there is a positive constraint between them; where they incohere, the constraint is negative.

"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, 3). Positive constraints can be satisfied either by keeping both of the locally cohering elements, or by rejecting them both. Negative constraints are satisfiable only by keeping one element and rejecting the other. The solution of a coherence problem gives a result that Thagard calls *global coherence*. The goal in various sorts of rational deliberation, according to Thagard, is to *maximize* global coherence—to pick the solution that satisfies as many constraints as possible (and thus which respects all the relations of local coherence for the domain as far as possible).

As Guarini notes, this coherence approach to modelling reasoning is what Thagard calls *biscriptive*, meaning that it is neither purely descriptive of how people actually do reason, nor purely prescriptive of how they should. As Thagard puts it, his model "describes how people make inferences when they are in accord with the best practices compatible with their cognitive capacities" (Thagard 1992, 97).

Thagard's biscriptive model of *ethical* reasoning, as Guarini tells us, actually depends upon a kind of *meta*-coherence—a coherence of at least four kinds of global coherence for four connected domains: explanatory, deductive, deliberative and analogical. Roughly, explanatory coherence generates the best fit with empirical data; deductive coherence produces the best fit between general principles and particular judgements; deliberative coherence gets you the best fit between judgements and goals; and analogical coherence brings out the best fit between judgements of some cases with judgements of other similar cases. These four sets of elements are, according to Thagard, connected by positive and negative constraints, and meta-global coherence is achieved by satisfying as many of these constraints as possible. Thagard calls this a *multicoherence* theory of ethics (Thagard 1998). This ethical theory is descriptive, insofar as it purports to model an idealized version of the way in which people actually make ethical

judgements, when they are being rational, but prescriptive insofar as it sets out to be an account of the way in which rational ethical judgements *should* optimally be made: the appropriate ethical judgement, presumably, is the one that maximizes the meta-global coherence of the ethical domain in question.

Guarini raises various problems with Thagard's account of ethical reasoning; these problems fall into three groups. The first set of problems Guarini broaches concerns the issue of determining the positive and negative constraints both between the elements of each domain (explanatory, deductive, deliberative and analogical), and between the domains themselves. Thagard does not (as far as I know) provide much detailed discussion on how these parameters are to be set in ethical deliberations,[1] but implicit in his discussion seems to be the assumption that the constraints will be to a large degree determined in such a way that they produce the correct answer for problems to which we already feel we know the answer (such as historical cases of scientific theory choice). If the network, so defined, produces good answers to a range of uncontroversial cases, it may be that we can then suppose that it will produce reliable answers to new problems. It could be, Guarini admits, that such a procedure will work for the making of scientific judgements "where there is quite a bit of agreement regarding which theories should have won when theories competed for acceptance" (Guarini 2003, 9) but it will not do in the sphere of ethics, where not only is there is no comparable pool of uncontroversial moral judgements, but there is not even agreement on how different kinds of consideration—such as empirical versus theoretical considerations—should be appropriately weighted with respect to each other. As a result, Guarini claims, Thagard's theory falls short of its claim to produce ethical judgements that can command general rational assent.

It seems to me, however, that the considerations Guarini is raising here point to two different sorts of problem with Thagard's model, and that, though Guarini does speak of two different 'underdetermination problems,' they cross-cut the distinction Guarini is trying to draw and as such are not very clearly distinguished. Furthermore, neither of the problems has much in the end to do with the lack of a suitable pool of uncontroversial moral cases for training purposes. The first sort of problem seems to me to be a *general* problem with programming a network in this *post facto* kind of way, and it is what Guarini is getting at when he says that "different assignments of weights for the different types of coherence [or, one might add, weightings *between* the 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" (Guarini 2003, 11). That is, in general, and presumably in the scientific sphere as in the ethical, non-equivalent networks could produce the same set of answers for any given finite collection of problems but produce different answers to future problems. We are thus faced with the problem of rationally deciding which of the two networks has the 'right' weighting. And notice that this problem cannot itself be solved by an appeal to coherence (at least, not at the level of the networks themselves) since the setting of the constraints for the problem domain is something that happens prior to, and sets the initial conditions for, considerations of global coherence.

The second sort of 'underdetermination' problem for Thagard that Guarini might be meaning to identify is of a quite different sort, and threatens to introduce precisely the sort of *circularity* to ethical deliberation that Thagard is keen to avoid. This is the concern that the setting of positive and negative constraints for a problem domain—which, recall, occurs prior to and independently of the determination of global coherence—is itself an ethical judgement. In particular, in determining the relative weighting of the four different kinds of coherence we are not relying merely upon such ethically-neutral considerations as deduction and association, but also on, for

example, our allegiances to particular ethical frameworks, such as consequentialism or deontology. Hence Thagard's hope that "once discussion establishes a common set of constraints, coherence algorithms can yield [moral] consensus" (Thagard 1998, 418) does not amount to the claim that *non*-moral foundational assumptions can be found which will pre-empt the need for ethical foundations (and in this Thagard's multicoherence account resembles other contemporary appeals to reflective equilibrium, such as, say, Rawls' 1996).

Guarini's next objection to Thagard seems to me to carry less weight. He argues, in essence, that Thagard is committed to the implausible view that "more entailments yield greater … coherence" (Guarini 2003, 12), but it is not immediately clear that this is true. Thagard is certainly committed to the position that *satisfying more constraints*, both positive and negative, yields greater global coherence, but it is not immediately clear how *this* commitment entails that "more positive linkage is better" (Guarini 2003, 12). Perhaps one way of capturing what Guarini is getting at is with the following example: consider a domain of elements together with their appropriate constraints. Imagine that some sorting of this domain into accepted and rejected elements produces the best possible degree of satisfaction of these constraints. Then suppose that, for example, we add an additional element that is entailed by some set of elements of the domain. This would add at least one additional satisfied constraint, and thus the new domain, and the new problem-solution, would apparently be *more coherent* than the earlier one.

All of this seems right, as far as it goes, but it is perhaps unnecessarily uncharitable to Thagard. Thagard's real position, surely, is that *for a given domain* maximizing global coherence consists in maximizing the number of constraints satisfied; there is no need to attribute to him the position that *larger* domains are more coherent than smaller ones, in virtue of their larger number of satisfied constraints.[2]

Guarini's third main objection is that Thagard's model is too computationally unwieldy to ever actually be self-consciously carried out by normal human reasoners, and that as such it cannot be a model of "how we reason when we reason at our best," as Thagard claims. Thagard considers a point very much like this one, and argues in detail that coherence problems run on connectionist networks are computationally tractable (Thagard and Verbeurgt 1998, 9–12). However, although Guarini concedes that it is possible that "some neural net in the brain settles on the belief set that has greatest coherence" (Guarini 2003, 14), he contends that this is not enough to account for the prescriptive force of moral reasoning: in order to *argue* that one moral position is superior to another, he suggests, one must be in a position to *know* that one's view is more globally coherent than the alternative, and it is *this* access which is denied us by our cognitive limitations—although our interpretation may in fact be maximally coherent, we cannot run the computation *in consciousness* in order to discern that it is.

While I think that Guarini's objection here has quite a lot of force, it bears pointing out that the worry seems to depend upon a kind of internalism about ethical knowledge—it suggests, I think, that we cannot know that ethical position *p* is justified without knowing that we know that *p*.[3] It might be open to Thagard to reply that, as long as we are reasoning well according to his biscriptive model, then we are rationally justified in our ethical beliefs, even if we cannot tell 'from the inside,' as it were, that we are justified. And in fact, Thagard not only stresses the piecemeal nature of our conscious moral reasoning, but also notes that ethical deliberation is an inherently *social* activity, and does not take place entirely within the confines of a single mind (Thagard 1998, 415).

Just as Guarini intends his objections to Thagard to be a friendly critique, so too do I intend my comments on Guarini in the same spirit. We both agree, I think, that Thagard's work on

computational models of moral reasoning is sophisticated and fruitful, and Guarini's critical assessment of Thagard is helpful and productive in moving the debate forward.

Notes

[1] Though he does discuss the constraints for *explanatory* coherence in some detail in Thagard 1989 and 1992.

[2] On a side-note, in his version of the Bernado case study Guarini says that positively constrained propositions "tend to be accepted together" and negatively constrained ones "tend to be rejected together": this is at odds with Thagard's actual account, which has positively constrained elements either jointly accepted or jointly rejected, and negatively constrained elements asymmetrically accepted or rejected.

[3] Of course this is quite a well-known problem with coherentism. See for example BonJour's struggles with the 'doxastic assumption' in his 1985.

References

BonJour, Laurence. *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

Guarini, Marcello. 2003. "Connectionist Coherence and Moral Reasoning." Unpublished manuscript.

Rawls, John. 1996. *Political Liberalism*. New York: Columbia University Press.

Thagard, Paul. 1989. "Explanatory Coherence," *Behavioral and Brain Sciences* 12: 435–467.

Thagard, Paul. 1992. *Conceptual Revolutions*. Princeton: Princeton University Press.

Thagard, Paul. 1998. "Ethical Coherence," *Philosophical Psychology* 11: 405–422.

Thagard, Paul, and Karsten Verbeurgt. 1998. "Coherence as Constraint Satisfaction," *Cognitive Science* 22: 1–24.