Brandom on the Sources of Normativity

Professor Joseph Heath
Département de philosophie
Université de Montréal

One of the most unsatisfactory sections of Robert Brandom's very complex and difficult book, *Making it Explicit*, is, unfortunately, the very first chapter.[1] Brandom's general objective in this work is to displace the concept of *representation* from its position as the central explanatory concept in the philosophy of language and epistemology, and replace it with some set of explanatory concepts derived from the analysis of social action or practice. In particular, he wants to argue that the concept of a social norm – a rule that determines, implicitly or explicitly, whether an action is correct or incorrect – can serve as a primitive concept in the development of a general theory of meaning. Successful execution of such a program would therefore constitute a vindication of some of the core intuitions underlying philosophical *pragmatism*.

The central problem with taking representation as an explanatory primitive, in Brandom's view, is that no one has ever been able to provide a satisfactory account of what the "representation" relation amounts to. Descartes, for instance, "notoriously fails to offer an account either of the nature of representational contents – of what the representingness of representations consists in – or of what it is to grasp or understand such content"(6). It is, indeed, difficult to explain representation without appealing to other, more sophisticated conceptual resources. In particular, representation seems to have an irreducibly normative dimension. Representation is not, for example, just causal covariance. To represent something is to represent it *correctly*. Once the misleading comparison to visual perception is set aside, representation therefore appears to be a very poor candidate for the role of explanatory primitive

---

[1] Robert Brandom, *Making it Explicit* (Cambridge, MA: Harvard University Press, 1994). Further page references appear in the text.

This critique of representationalism leads the reader quite naturally to expect that Brandom's preference for pragmatism is grounded in a sense that the explanatory primitives available within that framework are somehow less mysterious, or more intuitively accessible. In particular, one expects Brandom to show that the concept of "social norm" that he rests his analysis of language on can, in turn, be cashed out in terms of some simpler set of action-theoretic or behavioural concepts. And since preference for a pragmatic order of explanation is what *motivates* Brandom's whole project, it would not be unreasonable to expect an analysis of the action-theoretic primitives to appear front and centre at the beginning of the book. Furthermore, Brandom starts out in the first chapter *sounding* as if he is going to supply just such an analysis. Thus the absence of any conclusive argument or analysis on this score comes as something of a surprise. Many readers of *Making it Explicit* finish the first chapter not quite knowing whether Brandom chose to omit the argument, whether he made an argument, but a very weak one, or whether he chose to defer the burden of proof until chapter eight.

This confusion is lamentable. As Brandom notes, if one is willing to take the concept of representation as given, then there is a series of well-known mechanisms that can be used to generate from this an account of truth and inference, a theory of rational action, and so forth. Most of the detail in *Making it Explicit* consists in Brandom's attempt to show that, if one is willing to take the concept of a social norm as given, one can similarly generate an account of inference, truth, reference, and ultimately representation. However, the mechanism used to achieve the latter is still relatively unexplored, and fraught with technical difficulties. In order to make it worthwhile to iron out the kinks in this mechanism, Brandom must provide the reader with some reason to think that it is somehow more plausible to take normativity, rather than representation, as a primitive. The first chapter of *Making it Explicit*, however, manifestly fails to achieve this.

The goal of this paper is twofold. First, I will provide an analysis of the argument that Brandom does provide in chapter one, and develop an hypothesis intended to explain why that argument is left so inconclusive. My second goal is more forthright – I develop the account of the origins of normativity in social action that I think Brandom *should* have provided (indeed, *almost* provided).

I

Brandom's discussion of social norms is structured by his attempt to avoid two explanatory strategies that he considers to be unsuccessful. The first of these views, which Brandom calls *regulism*, identifies norms with some explicit formulation of a rule. Thus a social norm is understood on analogy with, say, a sign on the beach that says "no swimming." Normative assessment of action is possible because we can take a particular action, compare it against some rule that specifies how the action is to be performed, and determine whether it was done correctly or incorrectly. This view fails, according to Brandom, because the relevant species of normativity is merely subsidiary, or derived. "Proprieties of performance that are governed by explicit rules do not form an autonomous stratum of normative statuses, one that could exist though no other did"(20).

The problem with this type of regulism was clearly identified by Kant, but was given a more trenchant formulation by Wittgenstein. The basic objection takes the form of a regress argument. An explicit rule cannot, all by itself, determine the normative status of anything else. It must be applied. But application is itself something that can be done correctly or incorrectly. Thus there must be some second-order norm that specifies how the first-order norms are to be applied. But then how are these second-order norms to be applied? There must then be some third-order norm, that specifies how they are to be applied. A vicious regress ensues.

What is the lesson to be learned from this argument? According to Wittgenstein, "what this shows is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call 'obeying the rule' and 'going against it' in actual cases."[2] Or as Brandom puts it, "there is a need for a pragmatist conception of norms – a notion of primitive correctnesses of performance implicit in practice that precede and are presupposed by their explicit formulation in rules and principles"(21). The

---

[2] Ludwig Wittgenstein, Philosophical Investigations, trans. G.E.M. Anscombe (Oxford: Basil Blackwell, 1953), §201, p. 81e

regress arises because of the assumption that norms originally reside in "principles," and are applied to "practices" only secondarily. The problem is then that one can never get from principles to practices – more principles always seem to be required. The solution, therefore, is to formulate an account of what it means for there to be norms implicit in practices.

These considerations lead to a certain anti-intellectualism about norms. Wittgenstein's argument reminds us not to confuse the rules themselves with our explicit, or ideational, formulations of these rules. But while this reminder is helpful, it is possible to take this anti-intellectualism too far, and assume that the rules can be identified in a way that completely disregards our attitudes toward them. One such approach would be to identify social norms with simple regularities in conduct. This is the idea at the heart of the second type of explanatory strategy, which Brandom refers to as *regularism*.

The problem with regularism, according to Brandom, is that it loses sight of the distinction between what *is* done and what *ought* to be done. In other words, it loses sight of the properly normative dimension of social norms. One symptom of this difficulty – and a crucial objection to regularism – is the problem of gerrymandering. For any finite batch of behaviour, one can dream up an arbitrarily large number of rules of which that behaviour would be an instantation. As a result, when presented with a form of behaviour that appears to deviate from a rule, it is always possible to generate some other rule, with which that behaviour would be consistent.

The regularist's mistake is to think that just because norms do not consist in our explicit representations of them, our attitudes should be completely eliminated from the account. Thus the regularist hopes to discern the presence of norms simply by looking at behavioural regularities, while ignoring entirely the question of what agents take themselves to be doing. This is ultimately what generates the gerrymandering problem. Without some attention to what agents take themselves to be doing, there are simply too many rules that "fit" the data. Furthermore, everything is an instance of *some* rule. Thus the distinction between correct and incorrect performance collapses.

According to Brandom, "for the simple regularist's identification of impropriety with irregularity to get a grip, it must be supplemented with some way of picking out, as somehow *privileged*, some out of

all the regularities exhibited. To say this is to say that some regularities must be picked out as the ones that *ought* to be conformed to, some patterns as the ones that *ought* to be continued. The simple regularity view offers no suggestion as to how this might be done and therefore does not solve, but merely puts off, the question of how to understand the normative distinction between what is done and what ought to be done"(28).

Thus the trick is to find a way of acting that can be appropriately understood as an instance of "taking something to be correct" – and thus expresses the right sort of normative attitude – but which is not itself an explicit formulation of the idea that something or other is correct. Thus in order to find norms implicit in practice, we must first find normative *assessments* of action implicit in practice.

II

The most obvious candidate for a type of behaviour that manifests normative assessment is the *sanction*. We respond to actions that are correct with positive sanctions, actions that are incorrect with negative ones. A positive sanction can be understood here as anything that has positive gratificatory status for the agent acted upon, and hence reinforces the behaviour – a reward. A negative sanction is anything that has negative gratificatory significance, and hence conditions the agent not to repeat the behaviour – a punishment.

The most straightforward way of generating an account of social norms out of this conception of sanctioning is simply to define a norm as a sanctioned regularity in conduct. Brandom ascribes a theory of this sort to John Haugeland. According to such a view, agents conform to particular patterns because the pattern is positively sanctioned, or because any deviation from the pattern is negatively sanctioned, or both. Thus their actions are implicitly subject to normative assessment – an action is implicitly deemed to be correct when it is responded to with a positive sanction, and incorrect when it is responded to with a negative one. This sanction is what privileges a particular pattern, elevating it above the level of mere regularity.

There is something attractive about this account, since the sanction in question can be understood without presupposing other normative concepts, and yet clearly counts as a type of implicit normative assessment. Nevertheless, Brandom takes it to be inadequate. His central concern is that it is still a type of regularist theory, and so "merely puts off the issue of gerrymandering." The introduction of sanctions allows one to pick out a privileged pattern at the base level of behaviour. But the sanctioning itself is just another pattern of behaviour, and so can be understood as "enforcing" an arbitrary number of different rules. "Just as there is no such thing as *the* regularity of performance evinced by some actual course of conduct... so there is no such thing as *the* regularity that is being reinforced by a certain set of responses to responses, or even dispositions to respond to responses. The issue of gerrymandering, of how to privilege one specification of a regularity over equally qualified competitors, arises once more at the level of the reinforcing regularity"(36).

One way of putting the problem would be to say that, according to the simple sanctioning view, there is no way of telling whether the person doing the sanctioning is doing a "proper job" of it. The actions of the sanctioner are just a behaviour pattern, which can always be understood as an instance of some rule. Thus the distinction between "what ought to be done" and "what is done" disappears at this level.

One obvious way of trying to fix this would be to add on another level of sanctions – to treat the sanctioning behaviour as itself subject to further sanctions. This would be to recognize that "assessing, sanctioning, is itself something that can be done correctly or incorrectly"(36). This strategy, however, simply puts off the gerrymanding issue a bit further, and ultimately generates a regress. No matter how many "levels" of sanctioning are introduced, there will always be arbitrariness in the pattern at the "highest" level. Thus, according to Brandom, "if actual reinforcement of dispositional regularities is all that is available to appeal to in making sense of this regress, it might still be claimed that what is

instituted by this hierarchy of regularities of responses to regularities of responses ought not to count as genuinely normative"(36).[3]

(The account, incidentally, cannot be patched up just by turning to regularities of communal assessment either. Brandom observes that whether one person, or some group of people is involved in making the assessment, the gerrymandering problem persists.)

III

So what is Brandom's solution to all this? The answer, in his view, involves getting away from the idea that sanctioning needs to be understood in "naturalistic terms"(42). When sanctions are understood in terms of rewards and punishments, the goal is clearly to explain normative assessment in terms of some set of actions that can themselves be understood in nonnormative terms. But according to Brandom, "commitment to such a reduction is optional"(43).

Brandom acknowledges that one way to sanction someone is to do something that carries intrinsic gratificatory or deprivatory significance for that person. However, it can also count as a punishment to have one's normative status changed. Performing an action correctly might affect the range of actions that one is subsequently entitled to perform. Performing an action incorrectly might make it incorrect for one to attempt some further action. Thus the sanction that follows upon an action might be nothing more than a change in normative status. As Brandom puts it:

> As pointed out above, the assessing response constituting the community's acknowledgement of such a norm (the attitude corresponding to the status) might in some cases be describable in nonnormative terms – one who violates the norm is beaten with sticks, the norm-violating behavior is negatively reinforced. But other cases are possible, for instance ones in which the assessing response is to punish by making other actions inappropriate – one who violates the

[3] Brandom uses a peculiarly non-commital phrase here: "might still be claimed." The significance of this is unclear.

> norm is not permitted to attend the weekly festival. In such a case, the normative significance of transgression is itself specified in normative terms (of what is *appropriate*, of what the transgressor is *entitled* to do.)(43)

Brandom refers to a sanction that simply changes one's normative status as an "internal sanction" (since its *force* is internal to the normative system), and a sanction of the beating-with-sticks variety as an "external sanction."

Now clearly the introduction of internal sanctions will make absolutely no difference in the overall account if these sanctions just serve as proxies for external ones. One might argue that it is only the external sanctions that "really" count, and that each chain of internal sanctions has to be ultimately "anchored" in some sort of external ones. While one might not get beaten with sticks for, say, missing the hunt, just banned from the festival, one *will* be beaten with sticks if one subsequently tries to attend the festival. But then the whole thing can be redescribed in such a way as to avoid any reference to the internal sanction, simply by saying that anyone who misses the hunt *and* tries to attend the festival will be beaten with sticks. So this type of construction would be, in essence, no different from Haugeland's. The external sanction is still doing all the work, and so the problem of "gerrymandering" of the pattern underlying these sanctions persists.

Brandom's crucial departure from the Haugeland model, therefore, lies in his claim that internal sanctions *need not* ultimately be anchored in external ones (the restriction "can be relaxed," as he puts it). There is nothing incoherent, he argues, in a normative system, of which all the sanctions are internal. "Such an interpretation would not support any reduction of normative status to nonnormatively specifiable dispositions"(44). The system would be "norms all the way down"(44).

Unfortunately, Brandom does not say much more about this proposal. This is what is so baffling, given that his whole project hinges upon his being able to make some sense of the idea that there could be "norms implicit in practices." Brandom does not explain exactly how having a fully internalized set of sanctions moves us away from the regularist account. But more importantly, he does not say much to

allay the concern that a system of norms could not be sustained by internal sanctions alone. As soon as any external sanctions are introduced, however, there is the danger that the whole account will just dissolve back into Haugeland's.

The reason that one might be suspicious about the idea of a fully internalized system of sanctions is that it would only work under conditions of very strict compliance. In particular, the people being punished would always have to "play along" with their punishments. If a person broke one rule, he would not be directly corrected, but would simply lose some other entitlement that he had in the community. But this loss of entitlement would not be directly imposed, there would simply be a normative prescription that said "you're no longer entitled to do *x*." The agent would be free to disregard this rule, just as he disregarded the first. If the sanctions are all internal, the only way that others can respond to this second act of deviance is to impose another change in normative status, which the agent can once again choose to ignore. Without some type of external sanctions as a "last resort," no practice could persist in the fact of *resolute* deviance.

There is also a problem explaining how a community could ever induct new members into a set of practices, if the sanctions governing those practices were all internal. One would need to show that a practice which is "norms all the way down" is also learnable. It is a conspicuous feature of the way that we initiate children, for example, into our practices, that we rely quite heavily on sanctions to signal approval and disapproval. Often these sanctions involve cooperating, or witholding cooperation – parents are always saying things like "no, I won't pass it to you until you ask nicely for it." These are external sanctions. It is difficult to imagine that we could do without these sorts of sanctions entirely, or that the practices in which they are used are not "genuinely" normative on that account.

IV

Given these obvious objections to the idea of a practice governed entirely by internal sanctions, one is left wondering just how much Brandom thinks he has shown in this discussion. The best way of estimating this, I would argue, is to see that Brandom, in suggesting that a practice might be "norms all

the way down,' is not so much trying to provide an account of the action-theoretic underpinnings of normativity as he is offering a *principled refusal* to provide such an account.

The drive to "explain" normativity in terms of something more basic is, according to Brandom, guided by a desire (ultimately misplaced) to make norms "naturalistically respectable." Descriptive judgments, according to the typical naturalistic view, are not terribly mysterious. They are about the world. They are correct just in case the world is as they say it is. Prescriptive judgments, on the other hand – judgments that contain an *ought* in them – are much more mysterious. We don't really know what they are about, and we don't know how to decide whether they are correct or incorrect.

In order to make sense of norms, then, we have to provide some kind of a reduction. We have to explain norms in terms of some empirical, or "natural" phenomenon, such as behaviour. Thus what generates the dilemma for the regularist – whether the simple regularality theory, or the more sophisticated sanctions-based regularism – is its underlying "commitment to the possibility of a reduction of the normative to the dispositional"(46). When Brandom says that a practice may be "norms all the way down," he is suggesting that no such reduction may be possible, or necessary.

Brandom is comfortable with an anti-reductionist stance toward the normative because he thinks that the concept of the "natural" carries with it commitments that are ultimately just as mysterious as the concept of the normative. In particular, we take the "natural world" to a nexus of events linked together through *causal* relations. These causal relations, according to Brandom, can only be understood in terms of alethic modalities (necessity, possibility, impossibility). For example, it is part and parcel of causal relations that they support inferences to counterfactuals. Brandom also shares with Sellars the idea that these causal properties are built into the very concept of a wide variety of physical properties, and so he thinks that even simple descriptions carry with them commitment to claims that can only be expressed using modal vocabulary (103).

Thus the naturalist is, in effect, helping himself to the set of alethic modalities in formulating his claims, even though it is highly doubtful that these modalities can be cashed out in strictly naturalistic

terms. What Brandom is doing is therefore no worse. Instead of helping himself to the *alethic* modalities, he is helping himself to the *deontic* modalities (obligatory, permitted, forbidden – or something like that[4]).

Of course, this argument shows only the Brandom is no worse than the naturalist. But what could entitle him to take on these modalities in the first place? The answer to this lies in the recognition that the concept of normativity, along with the deontic modalities, are all pieces of *expressive* vocabulary. They are words that we language-using creatures introduce in order to talk about, in part, the practices that constitute our linguistic competencies. Both Brandom and the naturalist are introducing rival forms of expressive vocabulary. Whether one vocabulary is better than another is to be determined, not by its correspondence with the facts, but by the level of "expressive completeness"(641) it is able to achieve.

As a result, the question of whether we should take alethic or deontic modalities as primitive is to be determined by how good a story we are able to tell using one or another set of expressive resources. Brandom tries to show that, granted the deontic modalities, he is able to explain the rise of objectivity and representation, and to explain why our claims about the objective world support claims that are best formulated in terms of alethic modalities. The naturalist, on the other hand, even when helping herself explicitly to the alethic modalities, it still not able to give an adequate reconstruction of the deontic ones.

In the background of all this is Brandom's conviction that we already have an implicit grasp of what it is to follow a norm, because that is what we are doing whenever we make any sort of assertion. Thus he does not actually have to explain to the reader what it is to perform an act correctly or incorrectly. If the reader did not already understand this implicitly, she would be incapable of understanding the text. Thus Brandom's goal is expressive – it is to illuminate the structure of language from within. This is why he does not have to provide a "foundational" account of what norms are, or where they come from. So the general question that governs his inquiry, "where are the norms?" can be answered quite simply: "the norms turn out to be... here"(649).

---

[4] The qualification is there because Brandom takes entitlement and commitment to be the fundamental pieces of deontic vocabulary, not obligation and permission.

Ultimately, Brandom thinks that the prescriptive is more fundamental than the descriptive. Normativity is something that we grasp directly. We must have an implicit grasp of it to even get started in using language. The idea of a natural world, containing objects with properties, linked together by causal relations, is something that comes along much later. It is one of the highest achievements of our linguistic system – not a presupposition. Thus any philosophical system that attempts to make it an explanatory primitive is destined to remain deeply aporetic.

V

Followers of Brandom's work will of course notice that the above discussion brings to the foreground the "Hegelian" dimension of his philosophical views. One of the downsides of this Hegelianism is the general tendency to think that philosophical claims cannot be evaluated punctually, so to speak, but must be accepted and rejected at the level of whole "systems." Certainly one problem with Brandom's line on normativity is that his account of social norms stands and falls only with the "expressive completeness" of his entire theory. (This is also what generates that impression that Brandoms sees himself, in chapter one, as deferring some burden of proof).

One cannot help but wonder, however, whether Brandom's argumentation strategy doesn't contain an element of "sour grapes." After all, while Brandom may not take the physical world to be something unproblematically given, he shares with Sellars the view that something like conditioned responsive dispositions to one's environment can and should be taken as primitive (this is central to his account of observation). As a result, one cannot help but think that, were a simple behavioural account of what it is to follow a norm to become available, one that steered an acceptable middle course between regulism and regularism, Brandom would be more than happy to take it – and shift some emphasis away from the reliance on expressive completeness as an evaluative standard.

One strategy for developing such an account is suggested by the parallel that Brandom often draws between interpreting behaviour as *intentional* and interpreting practices as norm-governed. In adopting the "intentional stance" toward an organism, one is in effect choosing to use a certain

vocabulary in describing its actions. The core elements of this vocabulary are the concepts of *belief* and *desire*. Thus in adopting the intentional stance one is choosing to explain the organism's actions as goal-directed, and as guided by some set of representations pertaining to the achievement of this goal.

A theory that purports to explain intentionality in terms of the ascription of these states remains incomplete, however, insofar as it gives no account of the system that *does* the ascribing. One is inclined to think that any system which can adopt an intentional stance toward another must itself also be an intentional system.[5] Brandom therefore distinguishes between "simple" intentional systems and "interpreting" intentional systems. But where does this "interpreting" intentionality come from? If one posits some further intentional system, whose ascriptions constitute the intentionality of the interpreting system, then a regress has been initiated. What is needed, in order to resolve this regress in a satisfactory manner, is some kind of source for all this intentionality. What is needed is some type of *original* intentionality.

The most mechanical way of resolving this regress would be to abandon the "stance stance," and posit an intentional system whose intentionality is not inherited from some further system. John Searle uses such an argument to defend his view that some system must possess *intrinsic* intentionality. Brandom, however, wants to maintain allegiance to the stance stance. He rejects Searle's move, choosing instead to argue that original intentionality can found, roughly, in groups where agents each ascribe intentional states to one another. He states this claim as follows:

> The key to this account is that an interpretation of this sort must interpret community members as taking or treating each other in practice as adopting intentionally contentful commitments and other normatives statuses. If the practices attributed to the community by the theorist have the right structure, then according to that interpretation, the community members' practical attitudes institute normative statuses and confer intentional content on them; according to the

---

[5] In any case, Brandom thinks so. p. 58.

> interpretation, the intentional contentfulness of their states and performances is the product of their own activity, not that of the theorist interpreting that activity. Insofar as their intentionality is derivative – because the normative significance of their states is instituted by the attitudes adopted toward them – their intentionality derives from each other, not from outside the community. On this line, only communities, not individuals, can be interpreted as having original intentionality (61).

I would like to argue that this solution to the problem of the "origins of intentionality" provides a blueprint for a solution to the problem of the "origins of normativity." But first, a few words about how the solution to the intentionality problem works:

When we adopt the intentional stance toward some system, say a chess-playing computer, we ascribe to it a set of beliefs and desires. In other words, we try to anticipate its moves by assuming that it is "trying to win," and we don't worry about the nuts and bolts of its programming. In this case, the intentionality of the computer is clearly *derived*. It is only insofar as we treat it as having goals that it can be said to act in pursuit of them.

This interaction is characterized by a significant asymmetry, however. When I am playing against a computer, I am able to adopt the intentional stance toward it, but I assume that it is not able to do the same toward me.[6] What happens if the interaction becomes symmetric? Suppose the computer is replaced by an opponent whom I can reasonably regard as an interpreting intentional system. Once this substitution has been made, then I must treat my opponent not only as having beliefs and desires, but also as ascribing beliefs and desires to me. And if I think that this opponent is treating me not just as a simple intentional system, but also as an interpreting intentional system, then some of the beliefs that it

---

[6] As a matter of fact, this is why it is important to the success of these programs that chess is a finite decision problem – there is always in principle one best move. In any game with a significant strategic dimension – like

ascribes to me will be beliefs about its own beliefs, and beliefs about its beliefs about my beliefs, etc. Thus reciprocally imputed intentionality generates a regress of belief that increases the complexity of the interaction by several orders of magnitude.

One way of stating the difference that this increase in complexity imposes is to observe that the operations of simple intentional systems can be modelled decision-theoretically, while the operations of two interpreting intentional systems interacting with one another must be modelled game-theoretically. In a decision problem, the agent is able to form beliefs first, then select a utility-maximizing strategy. As a result, there will always be only one correct answer to any problem. In a game, agents cannot settle all their beliefs first, and so often can do no better than search for sets of beliefs and strategies that are consistent with one another. As a result, strategic choice problems can be indeterminate – they may have no single correct answer.[7]

Two points stand out here. First, there is clearly a sense in which the intentional states of these two systems are *instituted* by their adoption of the intentional stance toward each other. The regress of "interpreting" intentionality is closed off – everyone ascribes intentionality to everyone else, ascribes to everyone the ascription of intentionality to everyone else, and so on. This mutual ascription of interpreting intentionality is what constitutes original intentionality. The second point is that this account makes a certain sort of reductionism impossible. The mutual ascription of intentional states generates a characteristic indeterminacy in the behaviour of interpreting intentional systems when they interact with one another (formally recognized in the fact that strategic interactions – games – may have multiple equilibria). So while the adoption of the intentional stance toward a chess-playing computer is clearly optional, in the cases of two interpreting intentional systems interacting with one another, an observer

---

poker – computers have a lot more trouble, precisely because they are not able to model their opponent's mental states with any sophistication.

[7] For a more complete discussion of these issues, see Joseph Heath, "The Structure of Normative Control," *Law and Philosophy,* 17 (1998): 419-442.

would clearly be *missing something* if she failed to treat them as intentional systems. The fact that they are interpreting intentional systems is part of what makes them act the way that they do.

VI

The analogy between the intentionality problem and the normativity problem is fairly apparent. The general problem with regularity theories is that any segment of behaviour can be said to conform to an arbitrarily large number of rules. A norm must therefore be a pattern in this behaviour that is somehow "privileged." The question is, what is it that picks out one pattern, elevating it above all the others, making it the *right* thing to do, as opposed to just the thing that is done?

It is important to note that, in Brandom's view, sanctions do succeed in resolving this most basic problem. Once sanctions are applied to some segment of behaviour, it does have the effect of privileging one pattern. One can therefore be justified in claiming that there is a norm implicit in the initial pattern of behaviour, once it becomes sanctioned. The reason that the sanctioning solution is insufficient, in Brandom's view, is therefore not that sanctions are somehow inappropriate or insufficient to the task. The problem is that the application of sanctions generates a regress. It may become legitimate to describe some first agent's actions as correct or incorrect by virtue of some second agent's sanctions. But in order for there to be a norm, there must be some rule that the second agent's sanctions are enforcing. And since the second agent's sanctioning behaviour could be said to enforce an arbitrarily large number of rules, there would have to be something that privileges one of these patterns above all the rest. But if some third agent is introduced, whose sanctions will privilege some pattern in the second's, then a regress has been initiated.

So in the same way that an interpreting intentional system confers intentionality upon the simple intentional system (like the chess-playing computer), a system of sanctions confers normativity upon a pattern of behaviour. The regress begins when one asks where the interpreting intentional system gets its intentionality from, or where the sanctioning system gets its normativity from. Thus the Haugeland-style

"sanctioning solution" explains only the mechanism through which normativity is inherited. The source continues to elude us.

Following Searle's logic, one might be tempted to claim that this regress establishes the need for some pattern of behaviour that possesses "intrinsic normativity" – some pattern that is able to pick itself out as normatively privileged. (This is, for example, the strategy that Christine Korsgaard follows in her hunt for the sources of normativity.[8]) However, just as Brandom wants to maintain the "stance stance" in his discussion of intentionality, he also wants to maintain the idea that normative statuses are ultimately constituted by normative attitudes and assessments. Thus he is prevented from assuming that anything is inherently normative. His inclination, therefore, is simply to claim that the regress of normativity is harmless, or in any case irrelevant, since we already inhabit the space of the normative. This is what underlies his claim that our practices are "norms all the way down."

A much less problematic solution is suggested, however, by the approach that Brandom takes to the intentionality problem. Recall how the regress arises: Suppose one person acts. In order to say that this action is norm-conformative, we must introduce a second person, who will sanction the first. And in order to say that this sanctioning is norm-conformative, we must introduce a third person, who will sanction the second, and so on. Or so it would seem. But do we need to introduce the third person?

An easy way to eliminate the regress is just to close the circle after the first iteration. Instead of introducing a third person to sanction to second, we can simply stipulate that the *first* person sanctions the second. The second agent has what might be called an "expectation of behaviour" – she expects the first to behave in a certain way. If the first person anticipates these expectations, he may develop what we can call an "expectation of recognition" – he expects her to respond correctly to his actions, to punish him only when it is appropriate to do so, or to reward him when he is entitled to it.[9] Whenever either

---

[8] Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).

[9] This terminology, along with the general idea underlying the analysis, come from Jürgen Habermas, *The Theory of Communicative Action, Vol. 2* (Boston: Beacon Press, 1987), p. 19.

expectation is disappointed, sanctions are imposed. In this way, the second person's sanctioning efforts become subject to sanctions by the first, just as the actions of the first are subject to sanctions by the second.

Of course, this structure of reciprocal sanctions and expectations generates its own form of regress. The way that the first sanctions the sanctioning efforts of the second must also be sanctioned by the second. But this regress is clearly harmless in cases where all of these expectations and sanctions converge upon a single pattern of action. And these are precisely the cases in which would want to say that there is a norm implicit in the practice. When the second expects the first to do *x*, the first expects the second to expect the first to do *x*, the second expects the first to expect the second to expect the first to do *x*, etc., and all of these expectations are backed by sanctions, then only one action can satisfy all these expectations, viz. *x*. And so *x* is the *correct* action. Thus the regress, far from being vicious, generates something very much like the set of mutually reinforcing expectations that sustain game-theoretic equilibria.

As a result, when sanctioning is reciprocal, two agents can each act in a way that confers normativity upon the actions of the other, and this, by extension, confers normativity back upon their own actions. There is nothing left in the interaction that could count as "mere behaviour." In particular, because everyone engages in a normative assessment of everyone's conduct, everyone has no choice but to adopt such an assessment of their own conduct (at least implicitly). Thus it is plausible to suggest that "original normativity" inheres in the practices of a community in which everyone sanctions everyone else (and it does not matter whether these sanctions are "internal" or "external").

VII

This account of original normativity has two features that are attractive from Brandom's perspective (and are closely related). First, according to this view, normativity is not just a *façon de parler*. The vocabulary is not an optional component of our discussions of social interaction. Whenever we have a set of mutually sustaining normative expectations, the only way to properly understand the resulting

pattern of action will be in terms of these expectations. The second attractive feature of the account is that it makes normative interaction a sort of *sui generis* interaction structure, and hence satisfies Brandom's concern that one not entertain "any reduction of normative status to nonnormatively specifiable dispositions."

Let us take these two points in order.

When two individuals are interacting in a way that is genuinely normative, according to the account offered above, then an observer who comes along would clearly be missing something if he or she failed to interpret this conduct in normative terms. Consider again the analogy to game theory. Imagine an interaction between two players that has multiple Nash equilibria, and suppose that one of these equilibria is being played. This means that each player's action will be a best response to the action of the other. As a result, if we ask why player 1 chooses x, the answer will be that he expects player 2 to choose y. If we ask why player 2 chooses y, the answer will be that she expects player 1 to choose x. As a result, player 1's expectation that player 2 will choose y is grounded in player 2's expectation that player 1 will choose x, and vice versa. The fact that these two sets of expectations are mutually reinforcing is the only thing sustaining the behaviour pattern (since *ex hypothesi* neither action is intrinsically choiceworthy). Thus the only way to understand the interaction is to grasp this underlying set of beliefs and intentions.

Understanding a normatively regulated interaction imposes the same sort of demands. Imagine a simple rule specifying that x is normatively required of player 1. Such a rule can be instituted by a sanctioning structure that assigns player 2 some action, y, that is conditional upon player 1's performance of x. The set of normative expectations can then be closed by making player 1's obligation to perform x in subsequent iterations of the interaction conditional upon player 2's appropriate execution of y in response to x. It will then be the case that player 1 intends to perform x because it is the appropriate response to y, and that player 2 intends to perform y because it is the appropriate response to x. An observer who came along and didn't notice this underlying structure of normative expectations would again be missing something, since these expectations are constitutive of the pattern that the interaction exhibits.

The other major point worth noting is that neither player in a normatively regulated interaction needs to have any concrete disposition to perform any specific action. Player 1 may have no particular reason to choose *x*, other than the desire have player 2 choose *y*, and player 2 may have no reason to choose *y*, other than the desire to have player 1 choose *x*. As a result, it may not be possible to explain the interaction pattern in terms of any set of antecedently determined behavioural propensities, because such an account would fail to explain the crucial detail, viz. how *x* became the focus of these normative expectations in the first place.

The clearest illustration of this thesis can be seen in the fact that the emergence of genuinely normative interaction structures places limits on the range of human behaviour that can be explained using standard Darwinian evolutionary theory. The reason for this – to summarize a somewhat complex literature all-too-briefly – is that normatively regulated interactions create the possibility for cultural transmission of learned behaviour. This means that our behaviour, unlike that of other animals, must be explained using a "dual-inheritance" model, in which the contribution of genetic and cultural factors are explained using different models of transmission.[10]

All organisms have some ability to modify their phenotype in response to features of their environment. While some elements of the phenotype are genetically "hardwired," many more are learned. In its simplest form, learning requires only a mechanism that will make the animal more likely to perform actions that have been successful in the past. Thus two organisms with the same genotype may exhibit different phenotypes, depending upon aspects of the environment that affect their development.

Once such a learning mechanism is in place, it becomes possible to use sanctions in order to instill behavioural routines. This is called social learning. Such an ability allows the species to reliably reproduce certain phenotypic characteristics, even when they are not directly specified in the genotype. However, the phenotype can only stray so far, because the sanctioning behaviour, which ultimately

---

[10] Here I am drawing upon Robert Boyd and Peter J. Richerson, *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

generates the phenotypic effects, must still be specified in the genotype in order to be reliably reproduced.

However, once the sanctioning behaviour becomes itself subject to sanctions, it becomes possible for the phenotype to stray arbitrarily from the genotype. It becomes possible for what Boyd and Richerson call "population-level effects" to arise.[11] New phenotypes can arise and be "passed on," irrespective of the underlying genetic substructure. This is cultural evolution. Because this behaviour is not just transmitted "vertically" (from parent to child), but also "horizontally," (among sibs, peers, conspecifics), the selection pressures will function in a way that is quite different from the way that they operate in the genetic realm.

What does all this mean? Genetically determined behavioural propensities are sufficient to explain an enormous amount of animal behaviour, even very complex forms. Such propensities can also explain socializing behaviour, and so can indirectly explain elements of the phenotype that are acquired through social learning. However, once the socializing behaviour becomes itself the object of socialization (i.e. the sanctioning loop gets closed), this kind of explanation hits an in principle limit. Phenotypic forms can arise and be transmitted without any underlying genetic basis. Thus the normative becomes an irreducible component of our behavioural economy, and a source of *sui generis* phenotypic forms and population-level effects.

This point has extremely important consequences. There is an influential stream of pragmatism, perhaps formulated most influentially by Richard Rorty, which is inclined to think that the behavioural components of a pragmatist theory will ultimately be explained in terms of some theory of natural evolution. As Rorty puts it:

> What I retain is the conviction that Darwinism provides a useful vocabulary in which to formulate the pragmatist position... By 'Darwinism' I mean a story about humans as animals with

---

[11] Boyd and Richerson, *Culture and the Evolutionary Process*, p. 6.

> special organs and abilities: about how certain features of the human throat, hand and brain enabled humans to start developing increasingly complex social practices, by batting increasingly complex noises back and forth. According to this story, these organs and abilities, and the practices they made possible, have a lot to do with who we are and what we want, but they no more put us in a *representational* relation to an intrinsic nature of things than do the anteater's snout or the bower-bird's skill at weaving.[12]

If one subscribes to this "humans-as-slightly-more-complicated animals"[13] view, then there is no reason to assign normativity a very important position in the overall pragmatist order of explanation (at least insofar as one wants to deny that animals engage in genuinely normative practices). Thus Brandom is at pains to reject the idea that the difference between human and animal intelligence is just a matter of degree, and that our practices differ only in that they exhibit a higher level of complexity. (This is, for example, one of his reasons for distinguishing between intelligence and *sapience*.) A view like Rorty's is, from this perspective, ultimately a species of reductionism.

The account of normativity that I have outlined above is designed to provide a basis for Brandom's claim that there is a principled distinction to be drawn between rule-governed social practices and other forms of complex social behaviour. At the same time, it shows how a crude form of Darwinian reductionism – one which fails to recognize the distinctive contribution that the normative regulation of conduct makes to the structure of human social interaction – will be subject to explanatory and predictive deficiencies. Thus Brandom's goal of explaining normative statuses in terms of normative assessments is met, but without generating any vicious regress of normativity, and without running any risk of reductionism.

---

[12] Richard Rorty, "Putnam and the Relativist Menace," *Journal of Philosophy*. 90 (1993): 443-61, at 447-8.

[13] Ibid, 448.