VARIETIES OF SELF-REFERENCE

Brian Cantwell Smith

Intelligent Systems Laboratory, Xerox PARC
3333 Coyote Hill Road, Palo Alto, California 94304; and
Center for the Study of Language and Information
Stanford University, Stanford, California 94305

ABSTRACT

The significance of any system of explicit representation depends not only on the immediate properties of its representational structures, but also on two aspects of the attendant circumstances: implicit relations among, and processes defined over, those individual representations, and larger circumstances in the world in which the whole representational system is embedded. This relativity of representation to circumstance facilitates local inference, and enables representation to connect with action, but it also limits expressive power, blocks generalisation, and inhibits communication. Thus there seems to be an inherent tension between the effectiveness of located action and the detachment of general-purpose reasoning.

It is argued that various mechanisms of causally-connected self-reference enable a system to transcend the apparent tension, and partially escape the confines of circumstantial relativity. As well as examining self-reference in general, the paper shows how a variety of particular self-referential mechanisms — autonymy, introspection, and reflection — provide the means to overcome specific kinds of implicit relativity. These mechanisms are based on distinct notions of self: self as unity, self as complex system, self as independent agent. Their power derives from their ability to render explicit what would otherwise be implicit, and implicit what would otherwise be explicit, all the while maintaining causal connection between the two. Without this causal connection, a system would either be inexorably parochial, or else remain entirely disconnected from its subject matter. When appropriately connected, however, a self-referential system can move plastically back and forth between local effectiveness and detached generality.

1. INTRODUCTION

"If I had more time, I would write you more briefly." So, according to legend, said Cicero, thereby making reference to himself in three different ways at once. First, he quite explicitly referred to himself, in the sense of naming himself as part of his subject matter. Second, his sentence has content, or conveys information, only when understood "with reference to him" — specifically, with reference to the circumstances of his utterance. To see this, note that if I were to use the same sentence right now I would say something quite different (something, for example, that might lead you to wonder whether this paper might not have been shorter). Similarly, the pronoun 'you' picks someone out only relative to Cicero's speech act; the present tense aspect of 'had' gets at a time two millenia ago; and so on and so forth. Third, as well as referring to himself in these elementary ways, he also said something that reflected a certain understanding of himself and of his writing, enabling him to make a claim about how he would have behaved, had his circumstances been different.

In spite of all these self-directed properties, though, there's something universal about Cicero's statement, transcending what was particular to his situation. It is exactly this universality that has led the statement to survive. So we might say in summary that Cicero *referred to himself*, that the content of his statement was *self-relative*, that he expressed or manifested *self-understanding*, and yet that, in spite of all of these things, he managed to say something that didn't, ultimately, have much to do with himself at all.

Or we might like to say such things, if only we knew what those phrases meant. One problem is that thay all talk about the familiar, but not very well-understood, notion of 'self'. Perry [1983] has claimed that the self is so "burdened by the history of philosophy" as to almost have been abandoned by that tradition (though his own work, on which I will depend in the first two sections, is a notable exception). AI researchers, however, have rushed in with characteristic fearlessness and tackled self-reference head-on. AI's interest in the self isn't new: dreams of self-understanding systems have permeated the field since its earliest days. Only recently, however, has this general interest given way to specific analyses and proposals. Technical reports have begun to appear in what we can informally divide into three traditions. The first, which (following Moore) I'll call the *autoepistemic* tradition, has emerged as part of a more general investigation into reasoning about knowledge and belief (the theme of this conference). A second more procedural tradition, focusing on so-called meta-level reasoning and inference about control, is illustrated by such systems as FOL and 3-Lisp; for discussion I'll call this the *control* camp. Finally, in collaboration with the philosophical and linguistic communities, what I'll call the *circumstantial* tradition in AI has increasing come to recognise the pervasiveness of the self-relativity of thought and language (self-reference in the sense of "with reference to self").[1]

In spite of all this burgeoning activity, two problems haven't been adequately addressed. The first is obvious, though difficult: while many particular mechanisms have been proposed, no clear, single concept of the self has emerged, capable of unifying all the disparate efforts. Technical results in the three traditions overlap suprisingly little, for example, in spite of their apparently common concern. Nor has the general enterprise been properly located in the wider intellectual context. For example, as well as exploring the *self*, we should understand what sort of *reference* self-reference involves, and how it relates to reference more generally. Also, it hasn't been made clear how the inquiries just cited relate to the self-referential puzzles and paradoxes of logic (which, for discussion, I'll call *narrow* self-reference). At first glance the two seem rather different: AI is apparently concerned with reference to agents, not sentences, for starters, and with whole, complex selves, not individual utterances or even beliefs. We're interested in something like the lay, intuitive notion of "self" that we use in explaining someone's actions by saying that they lack self-knowledge. It isn't obvious that there is anything even circular, let alone paradoxical, about this familiar notion (folk psychology doesn't go into any infinite loops over it). And yet we will uncover important similarities having to do with limits.

The second problem is more pointed: there seems to be a contradiction lurking behind all this interest in self-reference. The real goal of AI, after all, is to design or understand systems that can reason about the *world*, not about *themselves*. Who cares, really, about a computer's sitting in the corner referring to itself? Like people, computers are presumably useful to the extent that they participate with us in our common environment: help us with finances, control medical systems, etc. Introspection, reflection, and self-reference may be intriguing and incestuous puzzles, but AI is a pragmatic enterprise. Somehow — in ways that no one has yet adequately explained — self-reference must have some connection with full participation in the world.

In this paper I will attempt to address both problems at once, claiming that the deep regularities underlying self-reference arise from necessary architectural aspects of any embedded system. Both cited problems arise from our failure to understand this — a failure attributable in part to our reliance on restricted semantical techniques, particularly techniques borrowed from traditional mathematical logic, that ignore circumstantial relativity. Once we can see what problem the self is "designed to solve", we'll be able to integrate the separate traditions, and explain the apparent contradiction.

The analysis will proceed in three parts. First, in section 2 I will assemble a framework in terms of which to understand both self and self-reference, motivated in part by the technical proposals just cited. The major insights of the circumstantial tradition will be particularly relevant here. Second, in section 3, I'll sketch a tentative analysis of the structure of the circumstantial relativity of any representational system. This specificity will be necessary in order to ground the third, more particular analysis, presented in section 4, of a spectrum of self-referential mechanisms. Starting with the simple indexical pronoun 'I', and with unique identifiers, I will examine assumptions underlying the autoepistemic tradition, moving finally to canvass various models of introspection and reflection that have developed within the control camp.

The way I will resolve the contradiction is actually quite simple. It is suggested by my inclusion of self-*relativity* right alongside genuine self-*reference*. Some readers (semanticists, especially) may suspect that this is a pun, or even a use/mention mistake. But in fact almost exactly the opposite is true — the two notions are intimately related, forming something of a complementary pair. Time and again we'll see how an increase in the latter enables a decrease in the former. For fundamental reasons of efficiency, all organisms must at the ground level be tremendously self-relative. On the other hand, although it enables action, this self-relativity inhibits cognitive expressiveness, proscribes communication, restricts awareness of higher level generalisations, and generally interferes with the agent's attaining a variety of otherwise desirable states. The role of self-reference is to compensate for this parochial self-relativity, while retaining the ability to act.

Explicit self-reference, that is, can provide an escape from implicit self-relativity.

Intuitively, it's easy to see why. Suppose, upon hearing a twig break in the woods, I shout "There's a bear on the right!" My meaning would be perfectly clear, but I have explicitly mentioned only one of the four arguments involved in the *to-the-right-of* relation;[2] the other three remain implicit and self-relative, determined by circumstance. However I can lessen the degree of implicit self-relativity by mentioning some of the other arguments explicitly. Look at this as a two stage process: one to get rid of the implicitness, one to get rid of the self-relativity (implicitness and self-relativity, that is, are distinct; *both* characterise ground-level action). In particular, the first move is to shift from the original statement to another that has roughly the same content, but that makes another argument explicit: "There's someone to the right *of me*." This latter statement is still self-relative, of course, but in a different, explicit, way. Now that I have a place for another argument, I can make the second move, and use a different expression to refer to someone else: "There's someone to the right *of you*", or "There's someone to the right *of us all*."

Thus the self provides a fulcrum, allowing a system to shift in and out of the particularities of its local situation. Both directions of mediation are necessary: neither totally local relativity, nor completely detached generality, would be adequate on its own. Roughly, the first would enable you to act, but thoughtlessly; the second, to think, but ineffectively.

So there is really no contradiction, after all. There is some irony, though: the self is the source of the problem, as well as being an ingredient in the solution. The overall goal in attaining detached general-purpose reasoning is to flush the self from the wings. However, the way to do that is first to drag it onto center stage. If you were to stop there, then you really would be stuck with a contradiction — or at least with a system so self-involved it couldn't reason about the world at all. Fortunately, however, once the self is brought into explicit view, it can then be summarily dismissed.

2. CIRCUMSTANCE, SELF, AND CAUSAL CONNECTION

2.A. ASSUMPTIONS

I'll focus on *representational* systems — without defining them, though I'll assume they include both people and computers, at least with respect to what we would intuitively call their linguistic, logical, or rational properties. For a variety of reasons I won't insist that representational systems be 'syntactic' or 'formal' (although what I have to say would equally well apply under what people take to be that conception).[3] Several other assumptions, however, will be important.

First, I take it that systems don't represent as indivisible wholes, in single representational acts, but in some sense have representational parts, each of which can be said to have content at least somewhat independently (*what* content a part has, however, will often depend on all the other parts — i.e., the parts don't need to be *semantically* independent). I take this notion of "part" very broadly: parts might be internal structures (tokens of mentalese, data structures, whatever), distinct utterances or discourse fragments issued over time, or even different aspects or dimensions of a complex mental state (what Perry has informally called mental "counties"). I will use 'agent' or 'system' to refer to a representational system as a whole, and 'representational structure' to refer to ingredients. When I specifically want to focus on the internal structures that are causally responsible for an agent's or system's actions, however, I will talk of *impressions* (as opposed to expressions, which I take to be tokens or utterances, external to an agent, in a consensual language). Impressions are meant to include data structures, elements of a knowledge representation system, or aspects of a total mental state. Such structures are sometimes classified abstractly (particularly in the "abstract data type" tradition), or identified with other abstract things to which they are thought to be isomorphic (like *beliefs*), but I will refer to them directly, because of my architectural bias and interest in causal role.

Second, representational structures are themselves likely to be compositionally constituted, which just means that they too may have parts (nothing is being said about compositional semantics, at least not yet). Again, the notion of part is rough; imagine something like a grammatical structure, or set of partially independent properties or elements, each of which contributes to the meaning of the whole. Utterances constituted of words according to the dictates of grammar are one example; composite structures in a data or 'knowledge' base are another. Thus the words 'I', 'would', 'have', and so on, are components of Cicero's claim (at least in its English translation). Since the term 'element' is biased towards ingredient objects and away from features or characteristics, and 'property' is biased the other way, I will refer to such parts as *aspects* of a structure or impression.

Finally, each constituent will be assumed to have what philosophers would call a *meaning*, which is something, probably abstract, that indicates just what and how it contributes to the content of the composite wholes in which it participates (i.e., I'm now adopting just about the weakest form of compositional semantics I can imagine). Meaning is not, typically, the same as content; rather, it's something that plays a role in giving a representation, or a use of a representation, whatever content it has. So the meaning of the word 'Katlyn' might be something like a relation between speakers and the world, a relation that enables those speakers, when they use the word, thereby to refer to whoever has that particular name in the overall situation being described. Though it's ultimately untenable, one can think of meaning as something a representational structure has, so to speak, *on its own*; the content arises only when it is used, in a full set of circumstances. So 'I' means the same thing when different people use it, but those uses have different contents.

As well as distinguishing meaning and content, we need to distinguish the latter — roughly, what a representation or statement is about — from an even more general notion of semantical *significance*, where the latter is taken to include not only the content but the full conceptual or functional role that the representational structure can play in and for the agent.[4] So for example in a computer implementation of a natural deduction system for traditional logic, a formula's content might be taken to be its standard (model-theoretic) interpretation, whereas its full significance would include its proof-theoretic role as well. It is distinctive of standard logical systems to view a sentence's meaning as the sole determiner of its content, and to take content as independent of any other aspect of significance. Situation theory [Barwise & Perry 1983] distinguishes meaning and content, and admits the dependence of the latter on circumstance, but takes both as specifiable independent of conceptual or functional role. In some of the cases we will look at, however, such as the use of inheritance mechanisms to implement default reasoning, all three will be inextricably intertwined.

2.B. CIRCUMSTANTIAL RELATIVITY

Given these distinctions, the most important observation for our purposes is that a great deal of the full significance of a representational system will not, in general, be directly or explicitly represented by any of the representational structures of which it is composed. Instead, it will be contributed by the attendant circumstances. Section 3 will be devoted to saying what "attendant circumstances" might mean, but some familiar examples will illustrate the basic intuition. As we've already seen, who the word 'I' refers to isn't indicated on the word itself, nor is it part of the word's meaning; rather, the meaning of 'I' is merely that it refers *to whoever says it.* Similarly, the referent of a pronoun may be determined by the structure and circumstances of the conversation in which it is used. If I say "solar tax credits have been extended for a year", the year in question, and the temporal constraints I place on it by using the past tense, emerge from the time of my utterance, not from anything explicit in the words. And, to take perhaps the ultimate example, whether what I say is *true* — which is, after all, part of its significance — is determined by the world, not (at least typically) by anything about the sentence itself.

Similarly, as the Carroll paradoxes show, the fundamental rules of inference can't themselves emerge in virtue of being explicitly represented, because further or deeper rules of inference would be required in order to use them. Nor do even the so-called "eternal" sentences of mathematics and logic carry all of their significance on their sleeve. That a predicate letter is a predicate letter, which is important to the interpretation of a formula in logic, is true in, but isn't represented by, that formula. Similarly, Lisp's being dynamically scoped isn't explicitly represented in Lisp. Or take the inheritance example suggested above: suppose you implement a representation system where a (representation of a) property attached to a node in a taxonomic lattice is taken to mean "an object of this type should be taken to have this property unless there is more specific evidence to the contrary". Thus, to use the standard example, if an impression of FLIES(X) is attached to the BIRD node, then the system is wired to believe that a particular bird will fly so long as there isn't an impression of ¬FLIES(X) attached in the lattice between the BIRD node and the individual node representing the bird in question. In such a system the content of the "so long as there isn't ... " part of the impression's meaning is architecturally determined: it is an implicit part of the overall system's structure, not explicitly represented, and it depends on the surrounding circumstances that obtain throughout the rest of the system, not on anything local to the particular structure under consideration.

This last example is intended to suggest why I am not distinguishing internal circumstance (whether there are other impressions standing in certain relational properties with a given one, say) and external circumstance (who is talking, where the agent is located, etc.). An informal division between the two will be introduced in section 3, but the similarities are more important than the differences, as evidenced in the similarities of mechanisms to cope with them. For one thing, since activity has to arise, ultimately, from the local interaction of parts, it may not matter whether a part's relational partner is somewhere across the system, or outside in the world; what will matter is that it's not right here. Also, the internal/external distinction isn't clean: since agents are part of the world in which they are embedded, some properties cross the boundary. For example, the passage of so-called 'real time' is often as crucial for internal mechanism as for overall agent.

2.C. EFFICIENCY

Before trying to carve circumstantial relativity into some coherent substructure, it's worth understanding why it's so pervasive. The answer has to do with efficiency, in a broad sense of that term. Specifically, in order for a finite agent to survive in an indefinitely variable world, it is important that multiple uses of its parts or aspects have different consequences, each appropriate to how the world is at that particular moment. Partly this enables a system to avoid drowning in details; any facts that are persistent across its experience can be "designed out", so to speak, and carried by the environment (as gravity carries the orientation argument for the human notion of to-the-right-of). But efficiency goes deeper, having also to do with how to cope with genuinely different situations.

The point is easiest to see in the case of action, where in fact it's so obvious as to be almost banal. Specifically, different occurences of what we take to be the same action have different consequences, depending on the circumstances of the world in which they take place. So if take a scoop with my back-hoe, what I get in the shovel will depend not on my action as such, but on the ground behind my tractor. Thus I can perfectly coherently say things like "after doing the same thing over and over, I suddenly cut the telephone cable." I.e., one can imagine viewing an action (read: meaning) as a relation between a local flexing of appendages and the situation in which that flexing takes place. The consequences of the action in a given situation (read: content) can be determined by applying the relation to the situation itself.

Action works this way because any other way of doing it would be horribly inefficient. Each day we want our actions to have different consequences (eating new meals, for example); it would be a terrible strain if we had to be structured differently for each one. As it is, we can have a finite and relatively stable structure, which can locally repeat doing the same things; the circumstantial relativity of perception and action will take care of providing the new consequences. The result is an efficient solution to what Perry characterises as a fundamental design problem:

> Imagine you want to populate the world with animals that will act effectively to meet their needs.
>
> There is one fundamental problem. Since these organisms will be scattered about in different locations, what they should do to meet their needs will depend on where they are and what things are like *around them.* This seems to present a problem. You can't just make them all the same, for you don't want them to do the same thing. You want those in front of nuts to lunge and gobble, and those who aren't to wander around until they are. (I have Grice's squarrels in mind.)
>
> You decide to make them each different. ... But then it strikes you that there is a more efficient way to do it. You can make them all the same, as long as you are a bit more abstract about it. You can make them all the same, [in the sense of having] their action controlling states depend on where they are. And you can do that, by giving them perception, as long as it is perception of the things about them. That is, you can make their internal states work in terms of what we have called subject relative conditions and abilities. You make them each go into state G when they are hungry and there are nuts in front of them, and each lunge forward and gobble when they are in state G.
>
> This way of solving a design problem, we call efficiency. [Perry 1983]

Like eating, representation needs to be efficient, and for similar reasons. First, actions are required in order to use and profit from the internal impressions: what page a least-recently-used virtual memory system discards, for example, will depend on circumstances. Second, impressions can themselves be circumstantially relative (what Perry calls "subject-relative") as both the pronoun and inheritance examples show. Finally, you would expect *ground-level* representations — representations connected directly with action and perception — to have the same (efficient) relativity as the actions and perceptions with which they are connected. Only in this way is there any hope of giving the connection between representation and action the requisite integrity. It is plausible to imagine a signal on the optic nerve directly engendering a rough impression of THERE'S-SOMETHING-TO-THE-RIGHT, but implausible to imagine its producing (and even this, of course, is still earth-relative):

RIGHT(SOMETHING, 38°N/120°W, 187°N, GRAVITY-NORMAL, 3-JAN-86/12:40:04)

Similarly, the stomach must first create the grounded impression "HUNGRY!"; it would take inference to turn this into "Won't you have some more pie?"

2.D. THE ROLE OF THE SELF

Circumstantial relativity isn't something an agent should expect to get over, but it has a down side. First, it doesn't lend itself to communication, if the relevant circumstances of the two communicators differ. If some agent A were simply to give agent B a copy of one of its representations, and B were to incorporate it bodily, the result might have completely different significance (and possibly even meaning) from the original. Information would not have been conveyed. If you're facing me, hear me say "there's a bear on the right!", take the sentence as your own, and then leap to *your* left, you would land in trouble.

Second, one of representation's great virtues is that it can empower a system with respect to situations remote in space or time, outside the system's own local circumstances. However, in order to represent those situations using impressions connected to those it uses to control action, the system must at least represent its own relativity, in order to be able to mediate between those less self-relative generalisations and more familiar implicit ones. I.e., to the extent that the content of its representational structures arise from implicit factors, it is impossible for a system to modify, discriminate with respect to, or make different use of any of the implicitly represented aspects of those representations' contents. If "HUNGRY!", without any argument, is the system's only means of representing the property of hunger, then it won't be able to represent any generalisation involving anyone else (such as that the bear on the right is hungry), or anything generic, such as that hunger sharpens the mind.

The third limit arising from circumstantial relativity depends on another fundamental fact about representation: its ability to represent situations in ways other than how they are. I will call this property of representation its partial *disconnection* (thus tree rings, under normal conditions of rainfall, don't quite qualify as representations because they are so nomically locked in to what they purportedly represent that they can't be wrong). A particular case of internal disconnection illustrates the third limit of circumstantial relativity. Typically, as long as some aspect of internal architecture isn't represented, the system will behave in the "standard" way with respect to that aspect. So, to consider the inheritance example again, the default FLIES(X) will always be interpreted by the underlying architecture in the "so long as there isn't ... " way. Suppose, however, that you want a variant on this behaviour: say, that the default should be over-ridden only if information to the contrary has been obtained from a reliable external source. Being implicit, however, the default way of doing things isn't available for this kind of modification. But if the internal dependence had been explicitly represented, then (as a consequence of the generative power of representation generally) the appropriate modified behaviour could probably be represented as well. In this way (under some constraints we'll get to in a moment) a system could alter its behaviour appropriately. In sum, explicit representation of circumstantial relativity paves the way for more flexible behaviour; without it, a system is locked into its primitive ways of doing things.

The representation of circumstantial relativity requires, among other things, the representation of one's self, because that self is the source of the relativity. There are of course different aspects of self, corresponding to different aspects of relativity: the self as a unity (such as for the *to-the-right-of* case), the self as a complex organization (for the inheritance example), the self as an agent (in generalising about the consequences of hunger).

Note that merely giving a system an impression that refers to it doesn't automatically solve the problem of circumstantial relativity. To see this, imagine installing within a system, as if by surgery, some impressions less self-relative than usual. For example, one might imagine providing a three-place representation RIGHT$_3$(X,Y,Z), and a distinguished token — say, $ME — to use as its own name. Chances are such representations would be conceptually possible, in the sense of not being architecturally precluded. They might enable an agent to reason (rather like a theorem-proving system) about some world. The problem would be that there would be no way for that system to act in that world, were it to find itself suddenly located there (no way for it to connect RIGHT$_3$ with the grounded THERE'S-SOMETHING-TO-THE-RIGHT!). The experience for the system might be a little like that of students who learn mathematics in a totally formal way (in the derogative sense), being able to manipulate formulae of various shapes around in prescribed ways, with no real sense of what they mean. Such a solution wouldn't make the representations *matter* to the system; they wouldn't connect with the agent's life. Furthermore, in a more realistic case where surgery is precluded (say, ours), there's no way to see how such representations could arise, given that they would have no direct tie to action or perception.

There's a problem, in other words: you've got to connect your explicit representations of circumstantial relativity with your grounded, circumstantially relative representations, which in turn connect with action. I will call this the problem of *appropriately connected detachment*. Entirely *disconnected* detachment, as the surgery example shows, is probably easy enough to obtain (at least in some sense), but it wouldn't be significant. Totally *connected* detachment is a bit of a contradiction in terms, but one can imagine an explicit representation so locked into the default circumstances that it wouldn't give you any power above and beyond what the grounded default case provided in the first place.

What is wanted is a mechanism that will continually mediate between the two kinds of representation — that will enable a system to shift, smoothly and flexibly, between indexical and implicit representations that can engender action, and generic and more explicit representations that enable it to communicate with others and in general have a certain detachment from its circumstances. The problem is to provide something like an ability to "translate" between the two kinds (or, rather, among elements arranged along a continuum, or even throughout a space — as we've seen, this is no simple dichotomy), just often enough to maintain the appropriate *causal connection* between located action and detached reasoning, but not so often as to lock them together. The right degree of causally connected self-reference, in other words, is our candidate for solving the problem of connected detachment. It enables a system to extricate itself from the limits of its its own indexicality, and yet at the very same moment to remain causally connected to its own ability to act.

There is one final thing to be said about self-reference mechanisms in general, before turning to particular varieties. In any representational system, the subject matter is represented in terms of what we might call a theory or conceptual scheme that identifies the salient objects, properties, relations, etc., in terms of which the terms and claims of the representation are stated. Except for some limiting simple cases, that is, representation is *theory-relative*. By this I don't mean so much relative to an explicit account, in the sense of a theory viewed as a set of sentences, but relative to a way of carving the world up, a way of finding oneself coherent, a scheme of individuation.

Granting this theory-relativity, we can see that causally connected self-reference requires the following three things:

1. A theory of the self, in terms of which the system's behaviour, structure, or significance can be found coherent. There is no *particular* aspect of the self that needs to be made explicit by this theory: we will see examples ranging from almost content-free sets of names, to complex accounts of internal properties and external relations.
2. An encoding of this theory within the system, so that representations or impressions formulated in its terms can play a causal role in guiding the behaviour of the system.
3. A mechanism of connection that enables smooth shifting back and forth between direct thinking about, and acting in, the world, and detached reasoning about one's self and one's embedding situation. The only example we have seen so far is a mechanism that mediates between K-ary representations of N-ary relations and K + 1-ary representations, as in the *to-the-right-of* case; more complex examples will emerge.

The first two alone aren't sufficient because they don't address the problem of causal connection. Thus the so-called 'meta-circular interpreters' of Lisp, as presented for example in [Steele & Sussman and 1978], meet the first two requirements, but there is no connection between them and the underlying system they are disconnected models of.

3. THE STRUCTURE OF CIRCUMSTANCE

I said earlier that particular mechanisms of self-reference can be understood as responses to different aspects of circumstantial relativity, which depend in turn on different aspects of circumstance itself. This means that, in order to understand these different mechanisms, we need an account of how circumstance is structured. This is a problem, for several reasons. First, there is probably no more problematic area of semantics. Second, we need a general account, since the whole point is to unify different proposals; nothing would be served by an account of how circumstance is treated by, say, semantic net impressions of a first-order language. Third, we *especially* can't assume the circumstantial structure of traditional first-order logic, since the whole attempt to make logical and mathematical language "eternal" can be viewed as an attempt to rid such systems of as much circumstantial relativity as possible. Although that goal isn't entirely met, as the Carroll paradoxes show, the formulae of logical systems certainly lack some of the important kinds of relativity that characterise embedded systems.

My strategy, given these difficulties, will be to give a rough sketch of the structure of circumstance. All I will ask of it is that it support the demands of the next section. Since my basic point is to show *how* the structure of self-reference reflects the structure of circumstantial relativity, any particular analysis of circumstance — including this one — can be taken as somewhat of an example.

By the *immediate* aspects or properties of a representational structure or impression I will mean those properties that can play a direct causal role in engendering any computational regimen defined over them. As such, they must not be relational — especially not to distal objects — but instead be locally and directly determinable, in such a way that a process interacting with or using the representation can "read off" the property without further ado (i.e., without inference). They must, that is, be immediately causally effective, in the sense that processes interacting with the structures can act differentially depending on their presence or absence.

For example, the (type) identity of tokens of a representational code (i.e., whether a given structure is a token of the word "elaborate" or not), how many elements a composite structure has, etc., would be counted as immediate. Non-immediate properties would include truth, being my favourite representation, and whether there is another type-identical representation elsewhere in a larger composite structure or system of which this particular representational structure is a part. This last example suggests that immediacy, which otherwise sounds like Fodor's notion of a *formal* property, is more locally restrictive, since all 'internal' properties of a computational system, it seems, count as formal to him.[5] Positive existence will count as immediate, but negative existence not, since there is nothing for the latter property to be an immediate property of.

Although it's tempting to compare the notion of an immediate property with apparently more familiar notions, such as of a syntactic, intrinsic, or non-relational property, such comparisons would involve us in more complexity than they're worth. The important point is merely that I mean to get at those aspects of a representational structure that affect or engender processes that use it; just what those properties *are*, especially in any given case, is less important.

In the last section I distinguished a system as a whole, its ingredient structures, and those structure's aspects or parts. With that set of distinctions, plus our semantic notions of meaning, content, and significance, plus the current notion of immediacy, we can define everything else we need. Specifically, I will say that something is *explicitly represented* by a structure or impression if it is represented by an immediate aspect of that structure. In contrast, something is *implicit* (with respect to an action or representation) if it is part of the circumstances that determine the content or significance of the representation or action, but is not explicitly represented. For example, I am explicitly represented by the sentence 'I am now writing section 3', since 'I' is a grammatical constituent of the sentence I use, and constituent identity is immediate. On the other hand, if I continue by saying 'but I should stop because it's after midnight', and the word 'midnight' represents the time in the Pacific Time Zone, then the Pacific Time Zone is an implicit part of the relevant circumstances. Similarly, if I say "there's a bear to the right", I am implicitly involved, but not explicitly represented.

There are shades of a use/mention distinction in the way I am characterising the implicit/explicit distinction: things are explicitly represented (nothing, yet, is explicit on its own) only if they are out there in the content, so to speak — part of the described situation, or referents. Something is explicitly represented, that is, only if it is mentioned, whereas something can be implicit either if it is used, or if it plays a middle role, not part of the sign itself, nor of the content or significance, but of the surrounding circumstance that mediates between the two. Thus the words of an utterance, on this view, are an implicit part of the circumstances that determine that utterance's content, since they are not themselves explicitly represented by the utterance (i.e, I am explicitly represented by 'I am writing', but 'I' plays only an implicit role). Where it won't cause confusion, however, I will also talk about explicit or implicit representations of things, as shorthand for representations that represent those things explicitly or implicitly.

Finally, by extension, I will say that something is explicit (*simpliciter*) only if it meets two criteria: it is explicitly represented, and it plays the role it plays in virtue of that explicit representation. So someone would be said to be an explicit part of a conversation only if they were explicitly refered to, and had whatever influence they had in virtue of that explicit representation. From this definition it follows that to *make something explicit* is to represent it explicitly in a causally connected way. Being implicit and explicit thus end up rather on a par, in the sense that both have to do with playing a role: to be *implicit* is to play a role directly; to be explicit is to play a role in virtue of being explicitly represented — which is to say, being represented by an immediate property.

We need to define one further notion, and then we are done. I have already called representational structures *self-relative* if different occurences of them (or things of which those occurences are a part) are part of the circumstances that determine their content. As pointed out above, however, there is more than one notion of part: part of the whole, and part of part of the whole. Rather than proliferating a raft of different notions of self-relativity, it will be convenient merely to separate the facts and situations of the overall circumstances into three broad categories: *external* circumstances, having to do with parts of the world in which the overall system is not a participant; *indexical* circumstances, including those situations in the world at large in which the system is a constituent, and *internal* circumstances, including both the ingredient impressions, processes defined over them, relations among them, etc. Thus who is President, and whether Shakespeare wrote the sonnet discovered in the Bodleian Library, would be paradigmatically external; where an agent was, and whom it was talking to, would be indexical. Internal circumstances would include whether a represented formula's negation is also represented; what inference rules can be, or are being, applied; how often this impression has been used since the system's last cup of coffee; etc. Finally, representations will derivatively be called *external*, *indexical*, or *internal* (or a mixture) depending on whether their content depends on the corresponding kind of circumstance.

This typology allows us to say all sorts of natural things: that the agent plays an implicit role in the significance of THERE'S-SOMETHING-TO-THE-RIGHT!; that 'I' is an explicit, indexical representation of an agent; that a truly unique identifier would be an explicit, non-indexical name; etc. Note also that a formula in a system of first order logic, at least in terms of its standard model-theoretic interpretation, has no implicit relativity to external or indexical circumstance (other than to the described situation itself), and no relativity to internal circumstance "outside" the formula, but aspects of it are nonetheless relative to the (implicit) internal structure of the formula itself (whether a variable is free, or what quantifier binds it, is implicitly determined by the structure of the expression containing it). Prolog impressions, however, are implicitly relative to internal circumstances of the beyond-formula variety (because of CUT, etc.), and are often used indexically. For example, the Prolog term RIGHT(JOHN,MARY), if it meant that Mary was to the right of John *from the system's perspective*, would be indexical.

4. VARIETIES OF SELF-REFERENCE

We can now show how various mechanisms of self-reference facilitate connected detachment.

4.A. AUTONYMY

I will call a system *autonymic* just in case it is capable of using a name for itself in a causally connected way. Just using a name that refers to itself doesn't make a system autonymic, even if that use affects the system in some way. What matters is that the name connect up, for the system, with its underlying, grounded, indexical architecture. To see this, imagine an expert system designed to diagnose possible hardware faults based on statistical analyses of reports of recoverable errors. Such a system might be given the data on its own recoverable errors, filed under a name known by its users to refer to it. The system's running this particular data set, furthermore, might eventually affect its very own existence (leading to board replacement, say). Even so, the system's behaviour wouldn't be any different in this case: it would yield up its conclusions entirely unaffected by the self-referential character of this externally provided name. When a system or agent reponds differentially, however — as for example do most electronic mail systems, which recognise and deal specially with messages addressed to their own users, forwarding other messages along to neighbouring machines — it will merit the label.

As we have already seen, two ingredients are required for autonymy. The first is a mechanism to convert K-ary impressions (of N-ary relations[6]) to K+1-ary impressions. For example, from the 0-ary HUNGRY! and unary RIGHT(SOMEONE), we need to produce HUNGRY(_), and RIGHT(SOMEONE,_). Second, we need a term, or name, to use so that the new, more explicit, version has the same content as the prior, implicit version. This is required because, on the story we're telling, it is this particular explicit version that, in virtue of being directly connected to the perceptual and action-engendering version, gives any more general versions their semantic integrity.

As the mail example suggests, something like a unique identifier can play this role. This is common in computational cases: designers of autonymic systems typically provide a way in which each system, though initially cast from the same mold, can be individually modified to react to its own unique name before being brought into service (a chore the system operators would do in "initialising" the system). As Perry suggests, however, this isn't efficient: it requires that each system be structured somewhat differently. What is distinctive about the pronoun 'I', in contrast, is that it gives exactly (type-)identical systems a way of explicitly referring to themselves. 'I', in other words, is an indexical term allowing explicit, but self-relative (hence efficient) self-reference. It doesn't on its own help a system to escape from its indexicality, but, because it makes that indexicality explicit, it is the minimal step away from fully implicit indexicality.

Causal connections to implement autonymy are so simple as to seem trivial, but their importance outstrips their simple structure. The mail systems provide a good example: that each mail host recognise its own name, and attach its own name to messages headed out into the external world, is a simple enough task, but crucial to the functioning of the electronic mail community.

4.B. INTROSPECTION

Purely autonymic mechanisms, in virtue of the inherent simplicity of names, are almost completely theory-neutral. By *introspective* systems, in contrast, I will refer to systems with causally connected self-referential mechanisms that render explicit, in some substantial way, some of their otherwise implicit internal structure. Since most of the self-referential mechanisms that have actually been proposed fall in this class, this variety of self-reference will occupy most of our remaining attention.

The first step, in analysing introspective systems, is to distinguish our own theoretical commitments from the theoretical commitments we attribute to the agents we study. The difference can be seen by comparing Levesque's [1984] logic of "explicit" and "implicit" belief (his terms, not ours, though the meanings are similar) with Fagin & Halpern's [1985] logics of belief and awareness. Levesque's use of B and L for explicit and implicit belief are predicates of the theorist; nothing in his account — as he himself notes — commits him to the view that the agents he describes parse the world in terms of anything like the belief predicate (i.e., in Fagin & Halpern's phrase, they need not be "aware" of the belief predicate). Fagin and Halpern, on the other hand, when they use such axioms as

$B\varphi \Rightarrow BB\varphi$, thereby commit the agents to an awareness of the same belief predicate they themselves use. I.e., for us to say "A believes φ" is for us to adopt the notion of belief; for us to say "A believes that it believes φ" commits A to the notion as well. Iterated epistemic axioms like $B\varphi \Rightarrow BB\varphi$ can therefore be misleading, since the inner *B*'s represents the agents' views; the outer ones the theorists'.

In the self-referential models typical of the autoepistemic tradition, the correspondence between explicit representation and belief is so close that this identification of agent's and theorist's commitment seems harmless, but when we deal with more complex introspective theories we will have to allocate theoretical commitments more carefully. For example, some theories that are straightforward, from a theorist's point of view, may be difficult or impossible for introspective systems to use, if they assume a perspective necessarily external to the agents they are theories of. Furthermore, different introspective theories require different primitive ("wired-in") support, whereas we, as external theorists, can use any theory we like, without fear of architectural consequence. For example, it is only a small move for a theorist to change from a theory of a programming languge that objectifies only the environment, to one that also objectifies the continuation. On the other hand, programming systems that can introspect using continuations are an order of magnitude more subtle than ones that introspect solely in terms of environments (we'll see why in a moment).

Keeping these cautions in mind, consider, as a first introspective example, an almost trivial autoepistemic computational agent comprising a set of base level representations, whose content, though perhaps self-relative, has primarily to do with facts about the world external to the system. As is usual in such cases, we will presume that the representation of each fact engenders the system's belief in that fact — we'll adopt, that is, the *Knowledge Representation Hypothesis* [Smith 1985] — so for familiarity we will call these representations *beliefs*, rather than impressions. Ignore reasoning entirely, for the moment, and assume that the agent believes only what has somehow been stored in its memory. For introspective capability, augment the base set of beliefs with a set of sentences formulated in terms of what Levesque calls an explicit belief predicate. So, for example, as well as containing the "belief" MARRIED(JOHN), imagine the system also being able to represent *B*(MARRIED(JOHN)).[7] We will call the whole system 𝔖, and its simple introspective representations *B*-sentences. (Note: In this and subsequent discussion I am representing impressions within 𝔖, not giving theoretical statements in a logic about 𝔖, so sentences of the form φ represent beliefs 𝔖 already has, and *B*-sentences represent introspective beliefs. All occurences of *B*, in other words, represent theoretical commitments on 𝔖's part.)

𝔖's *B*-sentences, though introspective, are still implicit and indexical, in several ways. First, the agent doing the believing — i.e., 𝔖 itself — remains implicitly (and efficiently) determined by internal circumstance, as does the current belief set with respect to which the *B*-sentence derives its truth conditions. I.e., $B\alpha$ is true just in case α is one of the base-level sentences, meaning that it is explicitly represented in 𝔖's general internal store, which will presumably change over time. Furthemore, by hypothesis, any implicitness or indexicality of 𝔖's base-level beliefs is inherited by the *B*-sentences: *B*(RIGHT(X)) is no more explicit about RIGHT's other three arguments than is the simpler RIGHT(X).

Given that 𝔖 is so simple, do the *B*-sentences do any useful work? Since we have claimed that introspective representations render explicit what was otherwise implicit, it is natural to wonder what otherwise implicit aspect of 𝔖's base-level beliefs these *B*-sentences represent. The answer requires a simple typology of "relations of structured correspondence". In particular, I will call a representation *iconic* (what is sometimes called *analogue*) if it represents each object, property, and relation in the represented domain with a corresponding object, property, and relation in the representation (iconic representations are thus fully explicit). Similarly, I'll say that a representation *objectifies* any property or relation that it represents with an object. Thus for example the sentence MARRIED(JOHN,MARY) objectifies marriage, since it uses (an instance of) the object 'MARRIED' to signify (an instance of) the relation of marriage that connects John and Mary. A representation *absorbs* any object, property, or relation that it represents with itself (thus the grammar rule EXP → OP(EXP,EXP) absorbs left-to-right adjacency). Finally, I will say that a representation is *polar* just in case it represents an absence with a presence, or vice versa (positive polarity in the first case, negative in the second). For example, the absence of a key in a hotel mail slot is often taken to signify the presence of the tenant in the hotel, making mail slots a negatively polar iconic representation of occupancy.

If all *B*-sentences were positive, then $\mathfrak{S}$'s introspective representations would be a partial, non-polar, iconic representation of its base level beliefs (partial because we're not necessarily assuming $B\alpha$ for all α). Since such representations objectify nothing, and therefore doesn't increase the explicitness of the base level, they aren't much use on their own. Causal connection for them is also obviously trivial. Negative *B*-sentences, however, of the form $\neg B\alpha$, make the introspective representations positively polar, thereby objectifying an otherwise implicit property of base level representations: namely, the property of negative existence (we have already seen that negative existence isn't immediate, which forces it to be implicit, unless explicitly represented as in this case). Thus $\neg B\alpha$ makes explicit one of the simplest imagineable implicit properties of a set of internal representations. No slight on importance is suggested, but it is noteworthy how close the correspondence between introspective impression and base-level impression remains: the objects of the introspective level correspond one-to-one with the objects of the base level; only a single, unary property is objectified (no relations); etc. Nonetheless, that one "rendering explicit" can have substantial computational consequences, because (once causal connection is solved) it makes immediate what wasn't otherwise immediate, with the effect that computational consequence can depend directly on the absence of a belief, which it couldn't do in the non-introspective version.

Causal connection, even with the positive polarity, is still relatively simple. $B\alpha$ will be true just in case α is an element of the set of representations, and although negative existence is not an immediate property of the belief set, constituent identity in a finite set is, so that it can be computed with only a moderate amount of inference — just a membership check on the base level belief set. Thus returning 'yes' or 'no' upon being *asked* $\neg B\alpha$ is relatively straightforward; it is perhaps less clear what should happen if $\neg B\alpha$ were *asserted*, although one can easily imagine a system in which this would either trigger a complaint, if α were already in the base set, or else perhaps cause its removal.

This example illustrates what will become an increasingly common theme: causal connection is typically easy or hard depending on two things:

1. The explicitness of the introspective representation (that is, the closeness of correspondence between the immediate properties of the introspective representation and its content); and
2. The immediacy of the aspects of self thereby explicitly represented.

An explicit representation of immediate properties of base-level beliefs, that is (which we have in this case), sustains relatively straightforward causal connection (this is really the point made in [Konolige

1985]). This equation — immediacy on both ends, simply connected — is hardly surprising, since immediacy is what engenders computational effect, and computational effect is required at both ends of causal connection. To the extent that immediacy on either end is lessened, or the connection becomes more complex, causal connection typically becomes that much more difficult.

Examples of such difficulty aren't hard to come by. They arise as soon as we complicate the example and consider introspective impressions that represent more complex internal properties — particularly relational ones. Curiously, in these more realistic cases introspective relativity itself tends to rise, as well as the non-immediacy of what is represented. Thus consider Moore's [1983] interpretation of $M\alpha$ as "α is consistent". This introspective representation is locally indexical because it is relative to the entire base-level set of representations, which isn't explicitly represented with its own parameter. Moore himself points out this relativity:

> "The operator *M* changes its meaning with context just as do indexical words in natural language, such as 'I', 'here', and 'now'. ... Whereas default reasoning is nonmonotonic because it is defeasible, autoepistemic reasoning is nonmonotonic because it is indexical."[8]

As it happens, however, this indexicality isn't what makes the causal connection of consistency difficult; rather, the problem stems from the fact that consistency itself isn't an immediate property, but a (computationally expensive) relational property of the entire base-level set. Similarly, when interpreted as "implied (or entailed) by the base level set", as in both Konolige and Fagin & Halpern, *B* is a relational, not immediate property (though again it is circumstantially relative), and causal connection consequently becomes problematic.

The environment and continuation aspects of the control structure of Lisp programs, made explicit in the introspective 3-Lisp, are also implicit, but not relational, and therefore more computationally tractable than consistency. 3-Lisp is so designed that causal connection is supported in both directions (see below); as well as obtaining a representation of what the continuation was, you can also cause the continuation to be as represented. So in 3-Lisp you can *assert* the introspective representation (whereas it is not clear what that would mean under the consistency reading of $M\alpha$, for example). Similarly, various different aspects of the Prolog proof procedure — goal set, control strategy, output — are made introspectively explicit in Bowen & Kowalski's amalgamated logic programming proposals. Again, the consistent assumption sets in a truth-maintenance system, typically implicit, are made explicit in deKleer's [1986] ATMS.

Since it would be hopeless to delve into these or any other introspective proposal in depth, I will devote the remainder of this section to three broad problems they all must deal with. First, however, it's important to note that the introspective models that typify the autoepistemic tradition represent an extremely constrained conception of introspective possibility. Admittedly, that tradition doesn't limit introspective beliefs to $B\alpha$ or $\neg B\alpha$, with *B* meaning "is immediately represented in the base level set", as our initial example suggests: the consistency reading of *M*, as Moore's example shows, and readings of *B* (or *L*) as "is implied by the rest of the belief set" are much more complex, as the discussion of causal connection makes clear. Nonetheless, such accounts can still largely be viewed as positively polar, iconic representations of derivable extensions of the base set. There is no inherent reason, however, to limit introspective deliberations to such one- or two-predicate vocabularies: one can easily imagine systems with introspective access to proof mechanisms and the state of proof procedures (as is typical in proposals from the control camp), or theories of self that deal with whether ground-level beliefs are chauvinist, creative, or largely derived from children's books. The kinds of meta-level reasoning that prompted AI's interest in self, cited for example in [Collins 1975], aren't limited to knowing *what* one believes, but having some understanding of it. The potential subject matter of introspection, in other words, is at least as broad as clinical psychology. In sum, whereas one can agree with Konolige's [1985] opening statement that "introspection is a general term covering the ability of an agent to reflect upon the workings of his own cognitive functions", there is no reason to limit those reflections as drastically as he does in constraining his "ideal introspective agents" to think nothing more interesting than "do I or don't I believe α?"

Introspective Integrity

The three issues that must be faced by any model of introspection are largely independent of basic cognitive architecture or theory of self. The first I call introspective *integrity*: it includes all questions of whether introspective representations are true, but extends as well to questions of whether any other significant properties they have (truth is only one) mesh appropriately with their content. In 𝕾's case integrity is relatively simple: $B\alpha$ should be represented just in case α is, and $\neg B\alpha$ just in case α is not. This simplicity depends partly on the simplicity of the introspective representational language, but also on another property of 𝕾 we haven't yet mentioned: the truth of 𝕾's introspective structures depends only on facts about the base-level representations, independent of introspective commentary. For an example where this doesn't hold, imagine a system where any impression (base-level or otherwise) is believed unless there is introspective annotation stating otherwise. Such a system would probably profit from an explicit representation of the truth and belief predicates, so that statements like "I should probably believe this, even though Mary doubts it", and "this can't be true, because it conflicts with something else I believe" could be represented (truth-maintenance systems are not unlike this). In such a case it would be natural to ask of any given base-level impression whether it is believed, but this can't be settled by inspecting only the base-level impressions. It would depend both on the state of the base level memory and on *implications* of the introspective commentary, and might therefore be arbitrarily difficult to decide. The truth-functional integrity of such a system would thus be inextricably relational.

Integrity is not offered as a property an introspective system must achieve, but rather as a notion with which to categorise and understand particular introspective axioms and mechanisms. For example, all of Konolige's notions of "ideality", "faithfulness", and "fulfillment" can be viewed as proposals for kinds of partial integrity. Similarly, Fagin and Halpern's $A_i\varphi \Rightarrow A_iA_i\varphi$ axiom for self-reflective systems is an axiom that ensures introspective integrity for their notion of awareness. In a particular case even outright introspective falsehoods could be licensed.

Truth isn't the only significant property, and therefore isn't the only aspect of integrity that matters, as we can see by looking at Bowen & Kowalski's DEMO predicate. According to the standard story, logic programs have both a declarative reading, under which clauses can be taken as formulae in a first-order language, and a procedural reading, under which they (implicitly) specify a particular control sequence, which implements a particular instance of the proof (derivability) relation. It follows that the *declarative* reading of DEMO should signify an abstraction over the (implicit) *procedural* regimen (i.e., $[\![\text{DEMO}]\!] = \vdash$, to be a little cavalier about notation). But this is not all that is required, if DEMO is to play the role they imagine: it must also be the case that the *procedural* reading of DEMO — i.e., the control sequence engendered by an instance of DEMO(PROG,GOALS) — must also lead to GOALS's being derived from PROG. Similarly, in 3-Lisp, where 'Φ' was used to signify content (i.e., roughly $[\![\; ... \;]\!]$), and 'Ψ' to indicate procedural consequence (roughly, $\vdash$), and where Ψ' (actually called NORMALISE) was the internal impression representing procedural consequence, it was necessary to show not only that $\Phi(\Psi') = \Psi$, but also, very roughly, that $\Psi(\Psi') \approx \Psi$. The general point is the following: suppose you have an impression A of some aspect P of the internal state (i.e., such that $[\![A]\!] = P$). In order for this to count as having *rendered P explicit* (rather than just as representing P explicitly), a use of this representation A of P must also *engender P* (remember, we said that something is rendered explicit only if it subsequently participates in the circumstances in virtue of that representation).

Intuitively, what this all means is that, in order to count as having introspective access to some aspect of your self, you must not only be able to represent that aspect, but you must be able to use that representation — step through it, so to speak, in what we informally call "problem-solving mode" — in such a way that this introspective deliberations *can serve as one way of doing what is being introspected about.* This might seem like a luxury, since after all there are things we can think about (such as how we ride a bicycle) that we can't simulate in virtue of reasoning with those thoughts. But one of the advertised powers of introspection is its ability to énable us to do things differently from how our underlying architecture would have done them. If we can't do them (introspectively) in the same way the architecture would have done them (non-introspectively), there seems little chance that we will be able to move beyond our base level capabilities. This is part of what causal connection demands. Thus, according to our account, although I can think about how I ride a bicycle, since I can't ride a bicycle by thinking about it, I can't call those thoughts causally-connected introspection.

Introspective Force

The second major issue, once again having to do with causal connection, is what I call introspective *force.* It has to do not with the causal efficacy of the introspective structures themselves, but with the causal connection between those structures and the aspects of self they represent. This is the problem addressed by what have been called 'linking rules', 'reflection principles', 'semantic attachment', 'level-shifting', etc.,[9] although simple quotation and disquotation operators are even simpler examples — e.g., InterLisp's KWOTE and (some of its uses of) EVAL; 3-Lisp's ↑ and ↓. In the discussion so far, I have characterised causal connection rather symmetrically, as a relation between representations and actual aspects of self. As the sophistication of introspection increases, however, the relation between self and self-representation not only grows more complex, but the two directions of connection — from self to representation (I'll call this "upwards"), and from representation to self ("downwards") — take on rather different properties. The differences are at least analagous to (what current ideology takes as) the distinction between beliefs and goals.

Imagine, to borrow an example from [Smith 1984], paddling a canoe through whitewater, exiting an eddy leaning upstream rather than downstream, and dunking. If, sitting on the bank a few moments later, you were to think about how to do better, you would first have to obtain an explicit representation of what you were doing just a moment earlier (this is the "belief" case: how do you go from a fact to a true belief about it?). It's no good to think "Ah, yes, the 20th century is drawing to a close"; you want to represent the very local situation that led you to fall into the river. This is the connection from reality (i.e., self) to representation. But similarly, after analysing the affair, and concluding that things would have gone better if you had leaned the other way, you don't want merely to sit on the bank, fatuously contemplating an improved self: the idea is to get back in the water and do better. You need, that is, a connection from representation to reality (more like what is called a "goal": you've got the representation; you want the facts to fit it). Both kinds of connection are germane even for as simple a self-referential representation as $\neg B\alpha$: the system might need to know whether $\neg B\alpha$ is true, or it might want to make it true. On $\mathfrak{S}$'s reading of B as "is explicitly represented" neither is too hard; if B means "consistent", the story, as we have already noted, would be very different.

As McDermott and Doyle [1980] discovered, it is easy to motivate perfectly determinate readings for introspective predicates where the causal connection isn't computable, even upwards. In the downwards case, moreover, if the property represented is a relational one, there may be no unique determinate solution (lots of things, typically, could make $\neg M\alpha$ true). It is thus a substantial problem, in actually designing an effective introspective architecture, to put in place sufficient mechanism to mediate between general introspectively represented goals and the specific actions on the self that have the dual properties of being causally connected (so that they can be put into effect) and satisfying the goal in question. This problem, however, is simply a particular case of the general issue of designing and planning action; since it isn't specific to the introspective case, it needn't concern us here.

Introspective Overlap

The third issue that must be faced by introspective systems is what I will call the problem of *introspective overlap*, which arises when the implicit circumstances of introspective impression coincide with, or include, what has been rendered explicit. The issue arises because the introspective representations are themselves part of what constitutes the agent. As such, any claims they make that involve, explicitly or implicitly, properties of the whole state of the agent, will be claims that they are likely, in virtue of their own existence or treatment, to affect. Introspective representations of relational properties that obtain between a particular impression and the whole set are obvious candidates for this difficulty. For example, if six beliefs were represented, one could not truthfully add the impression TOTAL-NUMBER-OF-BELIEFS(6); one would have to add TOTAL-NUMBER-OF-BELIEFS(7).

This overlap between content and circumstance is what opens the way for the puzzles and paradoxes of narrow self-reference. It is a more general notion than strict "circularity", since the problems can arise even if the representational structure itself is not part of its own content. An early but familiar example in computer science arose in the case of debugging systems for programming languages with substantial interpreter state, when written in the same language as the programs they were used to debug. These debugging systems, introspective by our account, rendered explicit the otherwise implicit parts of the control state of some other fragment of the overall system. The problem was that they too engendered control state (used global variables, occupied stack space, etc.), thereby introducing a variety of confusions because of unwanted conflict. These confusions often occasioned extraordinarily intricate code to sidestep the most serious problems, sometimes with only partial success. The fundamental problem, however, is easily described in our present terminology: the implicit aspect of the system that was rendered explicit remained the implicit aspect of the explicit rendering. There was no circularity involved, but there was overlap, with concommitant problems.

Overlap isn't necessarily a mistake: the indexicality that 'I' renders explicit is the same indexicality that implicitly gives the pronoun its content (similarly for 'here' and 'now'). Problems seem to arise only when negatives or activity affect what would otherwise be the case. It is typically necessary, in such cases, to give an introspective mechanism an appropriate *vantage point*, analogous to that provided by type hierarchies in logic, so that it can muck about with its subject matter without affecting the circumstances that make that subject matter its content.

Overlap only arises when the introspective machinery makes some implicit aspect of the internal circumstances explicit; it isn't a problem when what is implicit to the base-level is also implicit for the introspective machinery. Thus various systems, such as MRS and Soar, apparently don't make explicit any otherwise implicit state (everything that can be seen, self-referentially, is *already* explicit; what is implicit remains so), so the problem of overlap doesn't arise. In some other cases, such as in BROWN [Friedman and Wand 1984], overlap would occur, but the power of the introspective machinery is curtailed in advance to avoid contradiction. Handling overlap coherently was one of the problems that 3-Lisp was designed to solve: its purpose was to demonstrate the compatibility, in a theory-relative introspective procedural system, of detached vantage point, substantial implicit state, and complete causal connection (at the time I called 3-Lisp 'reflective', not 'introspective', but I now think this was a mistake: reflection — see below — was what I wanted; introspection was what I had). The continuation structures of 3-Lisp, representing the dynamic state of the overlapping processor, were what made it interesting. The other two aspects that were made explicit — structural identity, roughly, and lexical environment — didn't overlap (this is why, as we said earlier, an introspective variant of 3-Lisp that only rendered these two aspects explicit would be essentially trivial).

3-Lisp's particular solution to the problem of overlap was to provide what amounted to a type hierarchy for control, and in terms of that to provide, as a primitive part of the underlying architecture, mechanisms that always maintained the integrity of the connection between self-representation and facts thereby represented. So tight a connection was possible in 3-Lisp — because, as stated, continuations aren't relational — that it could be defined as equivalent (in an important sense) to the infinite idealisation in which all of its internal aspects (relative to its highly constrained theory) were always explicitly represented to itself. As a consequence, both external theorist and internal program could pretend, even with respect to recursively specified higher ranks of introspection, that it was indefinitely introspective with perfect causal connection. This particular architecture, however, clearly won't generalise to more comprehensive introspective theories, such as those involving consistency.

There is obviously no limit to the expressiveness of introspective representation, or intricacy of causal connection, though there are very real limits on the total combination of introspective expressiveness, integrity, and force. In the human case it seems clear that causal connection is the practical problem, especially in the "downwards" direction from representation to fact: though it's not exactly easy to come by psychological self-knowledge, it seems much harder, given such knowledge, to become the person you can so easily represent yourself to be.

The real challenge to self-reference, however, stems not from the limits on introspection, where after all one has, at least in some sense, access to everything being theorised about, but from the difficulty of obtaining a non-indexical representation of one's participation in the external world.

4.C. REFLECTION

In the last section a point was made that we need to go back to, because within it lie the seeds of the limits of introspective self-reference. In particular, it was pointed out, in connection with the move from the base-level RIGHT(X) to the introspective *B*(RIGHT(X)), that all of the implicitness of the former is inherited by the latter. The self-relativity of RIGHT — the fact that three of its four arguments get filled in by the indexical circumstances of the agent — is left implicit even in the introspective version. By a *reflective* system, in contrast, I will mean any system that is not only introspective, but that is also able to represent the external world, including its own self and circumstances, in such a way as to render explicit, among other things, the indexicality of its own embeddedness. This representational capacity, however, is (as usual) insufficient on its own; the system must at the same time retain causal connection between this detached representation, and its basic, indexical, non-explicit representations, which enable it to act in that external world.

Like substantial introspection, reflection is thus something we can only approximate; complete detachment is presumably impossible, both because no one knows to what extent properties that seem universal are in fact local but just happen to hold throughout our limited experience, and because it is very likely, for reasons of efficiency, that we won't ever have represented them. Reflection is also hard to attain, because of the requirement of causal connection. Finally, in order to obtain a representation of oneself that is truly external — i.e., that would hold from an external agent's perspective — one must first represent to oneself everything implicit about one's internal structure and state that isn't universally shared. Without this kind of self-knowledge, what one takes to be a detached representation of the world will still be implicitly self-relative, in ways one presumably won't realise. Introspection is therefore a prerequisite for substantial reflection (self-knowledge is a precursor of detachment). Yet in spite of these difficulties, reflection is necessary if one is to escape from the confines of self-relativity.

What then can we say about reflection, if it is so important? No very much, at least yet. Of the three self-referential traditions we've been tracking, neither the autoepistemic nor the control has addressed relativity to the external world. In both cases the self-referential focus has remained internal, though for different reasons. In the autoepistemic case, the "language" typically used for external representation either has either been, or been closely based on, mathematical logic, which, as Barwise and Perry have repeatedly emphasized, doesn't admit, in its foundations, of external relativity to circumstance. Hence logic's focus on sentences, rather than on statements, and its semantic models of mathematical structures, not situations in the world. In spite of all this, however, as pointed out earlier, even purely mathematical systems are permeated with internal implicitness: with questions of consistency, truth, etc. It is this internal relativity on which autoepistemic models of self-reference have therefore concentrated.

The control tradition stems more directly from computer science and programming language semantics, which have by and large trafficked in internal accounts. Its failure to deal with external relativity is roughly the dual of the autoepistemic's: whereas the autoepistemic tradition has dealt with external content, but not with relativity, computer science has focused on complex relativity, but not on the external world. Hence computer science's self-referential tradition — the control camp — has also dealt only with internal introspection. Programs, in particular, are typically viewed as (procedural) specifications of how a system should behave; as a result their subject matter is taken to be the internal world of the resulting system: its structures, operations, behaviour. Although one can (and I do) argue that the resulting computational systems are themselves representational, therefore bearing a "content" relation to the world in which they are ultimately deployed, that system–world relation isn't addressed by traditional programming language analyses. As a result, the implicitness represented by such self-referential models as metacircular interpreters [Steele & Sussman 1978], BROWN [Friedman and Wand 1984], MRS [Genesereth et al 1983], etc., is also primarily internal.[10]

Thus there is somewhat of a gap between the self-referential mechanisms that have so far been proposed (which are primarily introspective), and the accounts of external relativity offered by the circumstantial camp. What we need are mechanisms for rendering that external implicitness explicit. As usual, causal connection will be the difficult problem — more difficult than for introspection, since internal circumstance is always within the causal reach of the agent. The consistency of a set of first-order sentences may be difficult or impossible for a formal system to ascertain, but that isn't because there is crucial information somehow beyond the reach of that system, remote in time and space, to which other systems might have better access. Determining consistency is hard *all by itself.* The external circumstantial dependencies of ordinary language and thinking, however, are different: who is the right person to perform some particular function, for example, is something that only the world can ever know for sure. The best reflective agent will have direct causal access — and probably only partial access at that — to only one potential candidate.

This doesn't mean that serious reflection is impossible, however, partly because of our three-way, rather than two-way, categorisation of circumstance into external, indexical, and internal. The truth of whether Shakespeare wrote the sonnet is external; the implicitness motivated by efficiency, however, is typically indexical, not external, and indexicality has to do with the circumstances in which the agent participates — with circumstances, some of which, at least, should be relatively *nearby.* If there is any locality in this world, there seems more hope of an agent's knowing about local circumstances than about situations arbitrarily remote in space and time. What's enduringly difficult, of course, is that even those circumstances must be represented as if by another.

5. THE LIMITS OF SELF REFERENCE

Perfect self-knowledge is obviously impossible, for at least three reasons: because of the complexity of the calculations involved (such as those illustrated by consistency); because of the theory-relativity (no theory can render *everything* explicit); and because some circumstantial relativity — particularly indexical and external — is simply beyond the causal reach of the agent. But there are other limits as well. An important one stems from the fact that it is, ultimately, the same self that one is representing, and as such certain possibilities are physically excluded. The self can never be viewed in its entirety, because there is no place to stand — no vantage point from which to look.

Another limit — more a danger than a constraint — was intimated at the outset: although introspection (and self-knowledge) is a prerequisite to substantial reflection, it remains true that the power of all of these mechanisms derives ultimately from their ability to support more general, more detached, more communicable reasoning. It is a danger, however, that a system, in climbing up out of its embedded position, will end up thinking solely about its self, rather than using its self to get outside itself. This would lead to a self-involved — ultimately autistic — sort of system, of no use whatsoever.

These limits notwithstanding, self-reference and self-understanding are important. One can look out, see three people around the table, and represent the situation with "there are four people at this dinner party". One may also notice, perhaps with only introspective capability, that one is repeating oneself. But then one goes on to observe that, by doing so, one is acting inappropriately: that from the other three's perspective one looks like a fool. And then — here's where causal connection gets its bite — as soon as one has achieved this detached view of the situation, this representation from the outside, one scurries back into the introspective state, replaces the designator of that fourth person with 'I', recognises its special self-referential role, collapses back down to the fully implicit structures that engender talking, cuts them off, and thereby shuts up.

That's almost as good as writing more briefly.

ACKNOWLEDGEMENTS

I am indebted to everyone involved in AI's inquiry into self-reference, for their participation in what I take to be a collaborative inquiry. Particular recognition goes to Jim des Rivières, fellow traveller in the upper reaches of 3-Lisp, and to all other members of the Knights of the Lambda Calculus. Finally, a special debt is owed to John Perry: although most of the analysis in section 4 precedes his influence, much of the framework in which it is presented here, especially the emphasis on indexical relativity, is due to him. My thanks.

NOTES

1. For examples of the autoepistemic tradition, see for example [Fagin & Halpern 1985], [Konolige 1985], [Levesque 1984], [Moore 1983], and [Perlis 1985]. For the control tradition, see [Batali 1983], [Bowen & Kowalski 1982], [Davis 1976], [Davis 1980], [de Kleer et al 1979], [des Rivières and Smith 1984], [Doyle 1980], [Friedman & Wand 1984], [Genesereth & Smith 1982], [Hayes 1973], [Laird & Newell 1983], [Laird et al. forthcoming], [Smith 1982], [Smith 1984], and [Weyhrauch 1980]. For the circumstantial tradition, see [Kaplan 1979], [Barwise & Perry 1983], [Perry 1985a], [Perry 1985b], [Perry forthcoming], and [Rosenschein 1985]. Finally, I should mention those who have studied self-reference in specific cognitive tasks: for example [Collins 1975] and [Lenat & Brown 1984].
2. The fourth is orientation. Even if you and I are in the same place, and if A is to the right of B from my point of view, A will nonetheless be to the left of B from your point of view, if you happen to be standing on your head. Gravity establishes such a universal orientation that we rarely need to make this circumstantially determined argument position explicit.
3. Primarily because I don't think the notion of 'formality', as applied to computation, is coherent. See [Smith forthcoming (a)].
4. The term "conceptual role" is associated with Harman; see [Harman 1982], and [Smith 1984] for a computational account treating both content and conceptual role simultaneously.
5. However immediacy can also be less restrictive, since I will countenance some semantic properties as immediate, such as direct quotation, small arithmetic properties exemplified by immediate structures, etc. See [Fodor 1980], and [Smith forthcoming (a)]
6. For reasons that will be obvious, I don't think there is ever any reason — or need — to presume there is a final "fact of the matter" regarding how many arguments relations *really* have (or even that relations, as opposed to representations of them, *have* an arity). What is needed (for example in a scientific account) is a representation that makes explicit enough of the arguments so as to be able to convey, as widely as possible, insight, understanding, truth, whatever. If the universe were in fact an ordered progression of big bangs, numbered 1–..., with the relevant forces proportional to 1/k in each case (i.e., we're currently in the second round), all the relations of physics would turn out to have another parameter. That would be ok.

7. Or, if you prefer, *B*('MARRIED(JOHN)'). For purposes of this paper I don't need to take a stand on the question of the semantic or syntactic nature of believe objects, which is fortunate, because I no longer think it is a well-formed question. See [Smith forthcoming (b)].
8. [Moore 1983] pp. 6–7. By 'meaning' he means what we call content, and by 'indexical' he means what we mean by 'internally relative', but his point of course is valid.
9. 'Linking rule' is used in [Bowen & Kowalski 1982], 'semantic attachment' in [Weyhrauch 1980], 'level-shifting' in [des Rivières and Smith 1984], 'reflection principles' in [Weyhrauch 1980].
10. Not realising this fully at the time, I didn't initially describe 3-Lisp [Smith 1982, 1984] in a way that was very accessible to the programming language community. 3-Lisp's semantical model, in particular, was based on a conception of computation where the subject matter of a program was taken to include not only the system whose behaviour was being engendered, but also the subject matter of the resulting system. I still believe that this is often how programming is *understood*, even if implicitly, by a large number of programmers; my analysis, however, would have been more accessible had this non-standard semantic conception been treated more explicitly. Ironically, however, in spite of this semantical orientation, the only "external" world 3-Lisp was able to deal with was that of pure (and simple) mathematics, so it didn't really live up to its own semantical mandate.

REFERENCES

Barwise, Jon, and Perry, John (1983): *Situations and Attitudes*, Cambridge: Bradford Books.

Batali, John (1983): "Computational Introspection", M.I.T. A.I. Laboratory Memo AIM–701, Cambridge Mass.

Bowen, Kenneth A., and Kowalski, Robert A. (1982): "Amalgamating Language and Metalanguage in Logic Programming", in *Logic Programming*, ed. K. L. Clark and S.-A Tarlund, New York: Academic Press.

Collins, A. M., Warnock, E., Aiello, N, and Miller, M (1975): "Reasoning from Incomplete Knowledge", in *Representation and Understanding*, Bobrow, D. G., and Collins, A., eds., New York: Academic Press.

Davis, Randall (1976): "Applications of meta level knowledge to the construction, maintenance, and use of large knowledge bases", Stanford AI Memo 283 (July 1976), reprinted in Davis, R., and Lenta, D. B. (eds.) *Knowledge-Based Systems in Artificial Intelligence*, New York: McGraw-Hill.

——— (1980): "Meta-Rules: Reasoning About Control", *Artificial Intelligence* **15**: 3, pp. 179–222.

de Kleer, John, Doyle, Jon, Steele, Guy L., and Sussman, Gerry J. (1979): "Explicit Control of Reasoning", in *Artificial Intelligence: An MIT Perspective*, ed. P. H. Winston and R. H. Brown, Cambridge: MIT Press.

de Kleer, Johan (1986): "An Assumption-Based TMS", *Artificial Intelligence*, to appear 1986.

des Rivières, James and Smith, Brian C. (1984): "The Implementation of Procedurally Reflective Langauges", *Proc. Conference on LISP and Functional Programming*, pp. 331–347, Austin Texas. Also available as Xerox PARC Intelligent Systems Laboratory Technical Report ISL-4, Palo Alto, California, 1984.

Doyle, Jon (1980): "A Model for Deliberation, Action, and Introspection", M.I.T. Artificial Intelligence Laboratory Memo AIM–TR–581, Cambridge Mass.

Fagin, Ronald, and Halpern, Joseph Y. (1985): "Belief, Awareness, and Limited Reasoning: Preliminary Report", *Proceedings of IJCAI-85*, pp. 491–501, Los Angeles, California.

Fodor, Jerry (1980): "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology", *The Behavioural and Brain Sciences*, **3**: 1, pp. 63–73. Reprinted in Fodor, J. *RePresentations*, Cambridge: Bradford 1981.

Friedman, Daniel P., and Wand, Mitchell (1984): "Reification: Reflection without Metaphysics", *Proc. Conference on LISP and Functional Programming*, pp. 348–355, Austin Texas.

Genesereth, Michael R., and Smith, David E. (1982): "Meta-Level Architecture", Stanford Heuristic Programming Project Technical Report HPP–81–6, version of December 1982, Stanford California.

Genesereth, Michael R., Greiner, Richard, and Smith, David E. (1983): "MRS – A Meta-Level Representation System", Stanford Heuristic Programming Project Technical Report HPP–83–27, Stanford California.

Halpern, Joseph Y., and Moses, Yoram (1985): "A Guide to the Modal Logics of Knowledge and Belief: Preliminary Draft", *Proceedings of IJCAI-85*, pp. 480–490, Los Angeles, California.

Harman, Gilbert (1982): "Conceptual Role Semantics", *Notre Dame Journal of Formal Logic*, **23**, pp. 242–256.

Hayes, Patrick J. (1973): "Computation and Deduction", Proceedings of 1973 Mathematical Foundations of Computer Science (MFCS) Symposium, Czechoslovakian Academy of Sciences.

Kaplan, David (1979): "On the Logic of Demonstratives", in *Perspectives in the Philosophy of Language*, ed. P. A. French, T. E. Uehling, Jr., and H. K. Wettstein, Minneaspolis, pp 383–412.

Konolige, Kurt (1985): "A Computational Theory of Belief Introspection", *Proceedings of IJCAI-85*, pp. 502–508, Los Angeles, California.

Kowalski, Robert. (1979): "Algorithm = Logic + Control", *CACM* **22**, pp. 424–436.

Laird, John E., and Newell, Allen (1983): "A Universal Weak Method: Summary of Results", in *Proceedings of IJCAI-83*, pp. 771–773, Karlsruhe, West Germany

Laird, John E., Newell, Allen, and Rosenbloom, Paul S. (forthcoming): "Soar: An Architecture for General Intelligence", forthcoming.

Lenat, Douglas B., and Brown, John Seely (1984): "Why AM and EURISKO Appear to Work", *Artificial Intelligence* **23**, pp. 269–294.

Levesque, Hector J. (1984): "A Logic of Implicit and Explicit Belief", *Proceedings of the AAAI-84 Conference*, pp. 198–202, Austin, Texas. A revised and expanded version available as FLAIR Technical Report 32, Fairchild Artificial Intelligence Laboratory, Palo Alto, California, 1984.

McDermott, Drew, and Doyle, Jon (1980): "Non-Monotonic Logic I", *Artificial Intelligence* **13**:1&2, pp. 41–72.

Moore, Robert C. (1983): "Semantical Considerations on Nonmonotonic Logic", Artificial Intelligence Center Technical Note 284, SRI International, Menlo Park, California.

Perlis, Donald (1985): "Languages with Self-Reference I: Foundations", *Artificial Intelligence* **25**, pp. 301–322.

Perry, John (1983): "Unburdening the Self", unpublished manuscript, presented at the Conference on Individualism, Center for the Humanities, Stanford University, Stanford California.

——— (1985a): "Self-Knowledge and Self-Representation", in *Proceedings of IJCAI-85*, pp. 1238–1242, Los Angeles, California.

——— (1985b): "Perception, Action, and the Structure of Believing", in Grandy & Warner, eds: *Philosophical Grounds of Rationality*, Oxford: Oxford University Press, pp. 330–359.

——— (forthcoming): "Thought Without Representation", to be presented at a Joint Symposium of the Mind Association and the Aristotelian Society, London, July 1986.

Rosenschein, Stanley J. (1985): "Formal Theories of Knowledge in AI and Robotics," in *Proceedings of Workshop on Intelligent Robots: Achievements and Issues*, David Nitzan, ed., SRI International, Menlo Park, California.

Smith, Brian Cantwell (1982): "Reflection and Semantics in a Procedural Language", M.I.T. Laboratory for Computer Science Technical Report MIT-TR-272.

——— (1984): "Reflection and Semantics in Lisp", Conference Record of 11th POPL, pp. 23–35, Salt Lake City, Utah. Also available as Xerox PARC Intelligent Systems Laboratory Technical Report ISL-5, Palo Alto, California, 1984.

——— (1985), "Prologue to 'Reflection and Semantics in a Procedural Language' ", reprinted in R. Brachman and H. Levesque, eds., *Readings in Knowledge Representation*, Los Altos, CA: Morgan Kaufman, pp. 31–39.

——— (forthcoming a): "Is Computation Formal?", Stanford University CSLI Technical Report.

——— (forthcoming b): "Categories of Correspondence", Stanford CSLI Technical Report.

Steele, Guy L. Jr, and Sussman, Gerry J. (1978): "The Art of the Interpreter, or, the Modularity Complex (Parts Zero, One and Two)", M.I.T. Artificial Intelligence Laboratory Memo No 453, Cambridge, Mass.

Weyhrauch, Richard W. (1980): "Prolegomena to a Theory of Mechanized Formal Reasoning", *Artificial Intelligence* **13**: 1&2, pp. 133–170.