6 Varieties of Self-Reference

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 A1 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 A2 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 millennia 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 differed.

In spite of all these self-directed properties, though, there is something universal about Cicero's statement as well, 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 did not, 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). Researchers in Artificial Intelligence (AI), however, have rushed in with characteristic fearlessness and tackled self-reference head-on. AI's interest in the self is not 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 will call the **autoepistemic** tradition, has emerged as part of a more general investigation into reasoning about knowledge and belief.† A second more procedural tradition, focusing on so-called meta-level reasoning and inference about control, is illustrated by such systems as FOL and 3Lisp:‡ for discussion I will call this the **control camp**. Finally, in collaboration with the philosophical and linguistic communities, what I will call the **circumstantial tradition** in AI has increasing come to recognize the pervasiveness of the self-relativity of thought and language (self-reference in the sense of "with reference to self").[1]

†This paper was originally presented at a conference addressing this theme, entitled *Theoretical Aspects of Reasoning about Knowledge* (Monterey, California, March 19–22, 1986).

‡Throughout this book I have removed the hyphen in the name of this dialect, using '3Lisp' instead of '3-Lisp'.

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) [ch. 5], Doyle (1980), Friedman and Wand (1984), Genesereth and Smith (1982), Hayes (1973), Laird and Newell (1983), Laird et al. (forthcoming), Smith (1982) [ch. 3], Smith (1984) [ch.

In spite of all this burgeoning activity, two problems have not 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 surprisingly 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 has not been made clear how the inquiries just cited relate to the self-referential puzzles and paradoxes of logic (which, for discussion, I will call **narrow self-reference**). At first glance the two seem rather different: AI is apparently concerned with reference to *agents*, not to sentences, for starters—and with whole, complex selves, not individual utterances or even beliefs. We are 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 is not obvious that there is anything even circular, let alone paradoxical, about this familiar notion (folk psychology does not go into any infinite loops over it). And yet we will uncover important similarities having to do with limits.

A3

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

. . .

4], and Weyhrauch (1980). For the circumstantial tradition, see Kaplan (1979), Barwise and 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).

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 [fundamentally] 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. A4

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 will be able to integrate the separate traditions, and explain the apparent contradiction. A5

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 will 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 §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* alongside genuine *self-reference*. Some readers (semanticists, espe-

cially) may suspect that this is a pun, or even a use/mention mistake. But in fact almost exactly the opposite is true. [It is a fundamental thesis underlying the present analysis that] the two notions are intimately related, forming something of a complementary pair. Time and again we will see how an increase in the latter (self-reference) enables a decrease in the former (self-relativity). 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 A6 [basic] 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, [it will be argued,] 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 is easy to see why. Suppose, upon hearing a twig break in the woods, I shout "There is 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* characterize ground-level action). In particular, the first move is to shift from the original statement to another

2. The fourth is [vertical] orientation. Even if you and I are in essentially the same place, and looking out in the same direction, 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 [final?] circumstantially determined argument position explicit.

that has roughly the same content, but that makes another argument explicit: "There is 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 is someone to the right *of you*," or "There is 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. But there is some irony: 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 could not reason about the world at all. Fortunately, however, once the self is brought into explicit view, it can then A7 be summarily dismissed.

2 Circumstance, Self, and Causal Connection

2a Assumptions

I will focus on representational systems—without defining them, though I will 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 will not 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]

3. [I set formality aside] primarily because, [in spite of prevailing consensus,] I do not think the notion is in fact coherently applicable to compu-

Several other assumptions, however, will be important.

First, I take it that systems do not 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 do not need to be *semantically* independent). I take A8
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 A9
will use 'agent' or 'system' to refer to a representational system as a whole, and 'representational structure' to refer to [such] 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 [or communicative] 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 [computer science's] "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, [as well as severally constituting a complex system or agent as a whole,] 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 ele-

. . .

tation. See [Smith forthcoming (a)].

ments, 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 A10
(i.e., I mean now to embrace just about the weakest form of compositional semantics I can imagine). Meaning [in this A11
sense] is not, typically, the same as content; rather, it is 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 'Caitlyn' 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 whomever has that particular name in the overall situation being described. Though it is ultimately untenable, one can think of meaning as something a representational structure has "on its own", so to speak [i.e., in the sense of being independent of context of use]; the content arises only when it is used, in a full set of cir- A12
cumstances. 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 wider notion of [general] 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 A13
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† 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 A14
to implement default reasoning, all three will be inextricably intertwined.

2b Circumstantial Relativity

Given these distinctions, the most important observation for my purposes here 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 have already seen, whom the word 'I' refers to is not 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 utter-

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.
†Barwise & Perry (1983).

ance, not from anything explicit in the [meaning of 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. A15

Similarly, as the Carroll paradoxes show, the fundamental rules of inference cannot 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 is true in, but is not represented by, that formula. Similarly, Lisp's being dynamically scoped is not A16 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 is not 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 (not meaning!) of the "so long as there is not…" 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 is not right "here." Perhaps more significantly, the internal/external distinction is far from 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. A17

2c Efficiency

Before trying to carve circumstantial relativity into some coherent substructure, it helps to understand why it is 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 it is in fact so obvious as to be almost banal. Specifically, different occurrences 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 I take a scoop with my backhoe, what I pick up in its 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 the tractor's 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.

Our conception of actions works in this way because any other way of "parsing" it would be devastatingly inefficient. Each day we want our actions to lead to different consequences (eating new meals, for example); it would be a terrible strain A18
if we had to be structured differently for each one. As it is, we can have [or use] 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 A19
what Perry characterizes 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."†

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-IS-SOMETHING-TO-THE-RIGHT, but implausible to imagine its producing (and even this, of course, is still earth-relative):

A19.5

A20

RIGHT(SOMETHING, 38°N/120°W, 187°N,
GRAVITY-NORMAL, 3-JAN-1986/12:40:04)

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

†Perry 1983; pp. ….

2d The Role of the Self

Circumstantial relativity is not something an agent should expect to get over, but it has a down side. First, it does not lend itself to communication, if the relevant circumstances of the two communicators differ. If some agent A were simply to give another agent B a copy of one of its representational impressions, 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. A21
If you are facing me, hear me say "There is 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. How- A22
ever, 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 will not 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, do A23

not quite qualify as representations, on this account, because they are so nomically locked in to what they purportedly represent that they cannot be wrong). A particular case of internal disconnection illustrates the third limit of circumstantial relativity.

Typically, as long as some aspect of its internal architecture is not represented, a 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 is not…" way. Suppose, however, that you want a variant on this behavior: say, that the default should be over-ridden not if any *specific information to the contrary is represented, but only if that more specific contrary* information has been obtained from a reliable external source. Being implicit, however, the default way of doing things is not 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 modification of the default behavior could likely be represented as well. In this way (under A24 some constraints we will get to in a moment) a system could alter its behavior appropriately.

In sum, explicit representation of circumstantial relativity paves the way for more flexible behavior; without it, a system is locked into its primitive ways of doing things.

Among other things, the representation of circumstantial relativity requires 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 (useful in such cases as TO-THE-RIGHT-OF), the self as a complex organization (applicable to the inheritance example), the self as an agent (relevant to generalising about the consequences of hunger).

Note that merely giving a system an impression that refers

to itself does not 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 giving a system: (i) a three-place representation of "to-the-right-of"—say, $\textsc{right}_3(x,y,z)$; and (ii) a distinguished token—say, $\textsc{\$me}$—to use as its own name. Chances are that the provision of such representations would be conceptually possible, in the sense of not being architecturally precluded. They might enable the system or agent to reason (rather like a theorem-proving system) about some world. The problem would be that, without additional machinery, there would be no way for that system to act in that world, were it to find itself suddenly located there—i.e., no way for it to connect [an occasioning of] $\textsc{right}_3$ with [an occasioning of] 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. Merely providing such explicitised representations, and tying them into the system's general reasoning abilities, does not in and of itself make such representations *matter* to the system; they would not thereby be connected with the agent's life. Furthermore, in a more realistic case where surgery is precluded (say, ours), there is no way to see how such representations could *arise*, given that they would have no direct tie to action or perception. A25

There is a problem, in other words: you have 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 likely to be easy enough to obtain (at least in some architectural sense), but on its own A26

would not 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 would not give you any power above and beyond what the grounded default case provided in the first place. A27

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 own 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 have 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 partially 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 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 [or task domain] must be 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 do not 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. A28

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 behavior, 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 behavior of the system.
3. A mechanism of **appropriate causal 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 circumstances. The only example we have seen so far is a mechanism that mediates between k-ary and $k+1$-ary representations of n-ary relations, as in the TO-THE-RIGHT-OF case; more complex examples will emerge.

The first two alone are not sufficient because they do not address the problem of causal connection. Thus the so-called "meta-circular interpreters" of List, as presented for example in Steele & Sussman (1978), meet the first two requirements, but since there is no connection between themand the underlying system they are disconnected models of, they fail to meet the third. As such, they fail to meet the criterion of being able to serve as appropriately causally connected self-reference.

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* cannot 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 has not entirely been met, as the Carroll paradoxes show, the formulae of logical systems certainly lack some of the important kinds of relativity that characterize embedded systems.

My strategy, given these difficulties, will be to give a rough sketch of [some of the possible] structure of circumstance. All that I will ask is that it support the demands of the next section. Since my basic aim 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 A29

interacting with or using the representation can "read off" [the presence or absence of an instantiation of] the property without further ado (i.e., without inference). Immediate aspects or properties, that is, must be immediately causally effective, in the sense that processes interacting with the structures can act differentially depending on their presence or absence—depending on whether or not they are occasioned.

For example, the (type) identity of tokens of a representational code (i.e., whether or not a given structure is a token of the word 'elaborate'), how many elements a composite structure has, etc., would on this account 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 is 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 are worth. The important point is merely that, with the notion of immediacy, I mean to get at those aspects of a representational structure that [are available to] affect or engender processes that use it; just what such potentially effective properties *are*, especially in any given case, is less important. A30

5. Immediacy can also be less restrictive than formality, however, since I will countenance some semantic properties as immediate, such as the reference of direct quotations, small arithmetic properties exemplified by immediate structures, etc. See Fodor (1980) and Smith «forthcoming (a)».

In the last section I distinguished a system as a whole, its ingredient structures, and those structure's aspects or parts. With (i) that set of distinctions, (ii) our semantic notions of meaning, content, and significance, and (iii) the current notion of immediacy, we have in hand everything we need [to lay out the account of self-reference].

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 of this paper," since 'I' is a grammatical constituent of that sentence, and constituent identity is immediate. On the other hand, if I continue by saying "but I should stop because it is 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 [even though it is not part of the reference of 'midnight'—i.e., of the metaphysical moment thereby referred to]. Similarly, if I say "There is 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 characterizing 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 A31
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 circumstances that mediate between the two. Thus the words of an utterance, on this view, are an implicit part of the circumstances that determine that utterance's con-

tent, since they are not themselves explicitly represented by the utterance (i.e., I am explicitly represented by the sentence "I am writing," but in that sentence the word 'I' plays only an implicit role). Where it will not 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: (i) it is explicitly represented, and (ii) 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 *referred 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. A32

We need to define one further notion, and then we are done. I have already called representational structures *self-relative* if different occurrences of them (or things of which those occurrences 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 mereological 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 im-

pressions, processes defined over them, relations among them, etc. Thus who is President, at the time of any given utterance or act of reasoning, and whether Shakespeare wrote the sonnet discovered in the Bodleian Library, would be paradigmatically external. Where a person or reasoning agent was, and whom it was talking to, would be (for it) 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-IS-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 an occurrence of variable is free, for example, 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 such operations as 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 counted as indexical.

4 Varieties of Self-Reference

We are now finally in a position to show how various mechanisms of self-reference facilitate various forms of connected detachment.

4a. Autonymy

I will call a system **autonymic** just in case it is capable of using a name for itself in an appropriately causally connected way. Just using a name that refers to itself does not 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 behavior in this case would not be any different from its behavior in any other; it would yield up its conclusions entirely unaffected by the self-referential character of this externally provided name. When a system or agent responds differentially, however as for example do most electronic mail systems, which recognize and deal specially with messages addressed to their own users, forwarding other messages along to neighbouring machines—it will merit the autonymical label.

As we have already seen, two ingredients are required for autonymy. The first is a mechanism to convert between k-ary and $k+1$-ary impressions of n-ary relations.[6] For example,

6. For reasons that will be obvious, I do not 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,

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 are 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 explicit 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 "initializing" the system). As Perry suggests, however, this is not 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. On its own it does not help a system 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 recognize its own name, and attach its own name to messages headed out into the external world, is a simple

. . .

insight, understanding, truth, whatever. If the universe were in fact an ordered progression of big bangs, numbered 1–…, with k spatial dimensions and forces proportional to $1/r^{k-1}$ in each case (i.e., we are currently in the third round), all the relations of physics would turn out to have another parameter. That would be OK.

enough task, but absolutely crucial to the functioning of the electronic mail community.

4b 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, A33 this variety of self-reference will occupy most of our remaining attention.

The first step, in analyzing 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 logic of "explicit" and "implicit" belief[†] (his terms, not ours, though the meanings are similar) with Fagin & Halpern's logics of belief and awareness.[‡] Levesque's use of the predicates 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 Bφ ⊃ BBφ, 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 of belief as well. Iterated epistemic axioms such as Bφ ⊃ BBφ can therefore be substantially misleading, since any inner (non-initial) B's must represent the agents' notion; the outer ones will be only the theorists'.

[†]Levesuqe (1984).
[‡]Fagin & Halpern (1985).

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 language 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 will see why this is so 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, within the system, engenders the system's belief in that fact—that is, we will adopt the *Knowledge Representation Hypothesis laid out in* 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

†Included here as §4 of ch. 3a, p. 8.

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] I 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 an external logic *about* S, so sentences of the form φ represent beliefs S already has, and B-sentences represent introspective beliefs. All occurrences 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(α) 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. Furthermore, 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, A34

7. Or, if you prefer, B('MARRIED(JOHN)'). For purposes of this paper I do not 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)».

and relation in the representation (iconic representations are thus fully explicit). Similarly, I will say that a representation A35
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. A36

If all B-sentences were positive, then S's introspective representations would be a partial, non-polar, iconic representation of its base level beliefs (partial because we are not necessarily assuming B(α) for all α). Since such representations objectify nothing, and therefore do not increase the explicitness of the base level, they are not of much use on their own. Causal connection for them is also relatively trivial. Negative B-sentences, however, of the form ¬B(α), 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 is not immediate, which forces it to be implicit, unless explicitly represented, as in this case). Thus ¬B(α) makes explicit one of the simplest imaginable 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, as logicians are not the only ones who know, that single act of "rendering something explicit" can have substantial computational consequences, because—once appropriate causal connection is provided—it makes immediate what was not otherwise immediate, with the effect that computational consequence can depend directly on the absence of a belief, which it could not (at least not easily) do in the non-introspective version.

Causal connection, even with the positive polarity, is still relatively simple. B(α) will be true just in case α is an element of the set of representational impressions, and although negative existence is not an immediate property of the belief set, constituent identity in a finite set is, so that negative existence 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* "B(α)*?*" is relatively straightforward. It is less clear what should happen if ¬B(α) were to be *asserted*, although one can easily imagine a system in which this would either trigger a complaint, if α were already in the base set of impressions, or else perhaps cause its removal.

This example illustrates what will become an increasingly common theme: whether 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. A37

An explicit representation of immediate properties of base-level beliefs, that is (such as their "syntactic" properties, their presence or absence, which we have in this case, etc.), sustains

relatively straightforward causal connection.[†] 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 either (i) immediacy on either end is lessened, or (ii) the connection between them becomes more complex, causal connection typically becomes that much more difficult.

Examples of such difficulty are not 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 $\mathrm{M}(\alpha)$ as "α is consistent." This introspective representation is locally indexical because it is relative to the entire base-level set of representations, which is not 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 is not what makes the causal connectivity of consistency difficult; rather, the problem stems from the fact that property of consistency is *not itself immediate*, 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

[†]This is really the point made in Konolige (1985).

8. Moore (1983) pp. 6–7. By 'meaning' Moore means what we are here calling content, and by 'indexical' he means what we mean by 'internally relative,' but his point of course is valid.

Konolige and Fagin & Halpern,[†] B becomes a relational, not immediate property (though again it is circumstantially relative), and causal connection consequently grows problematic.

The environment and continuation aspects of the control structure of Lisp programs, made explicit in the introspective 3Lisp,[‡] are also implicit, but not relational, and therefore more computationally tractable than consistency. 3Lisp 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 3Lisp you can *assert* the introspective representation (it is not clear what that would mean under the consistency reading of M(α), 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 assumption-based truth maintenance system ATMS.[**]

Since it would be hopeless to delve into these or other introspective proposals in depth, I will devote the remainder of this section to three broad problems they all must deal with. Before doing so, however, it is important to note that the introspective models that typify the autoepistemic tradition represent an extremely constrained conception of introspective possibility. Admittedly, that tradition does not limit introspective beliefs to B(α) or ¬B(α), 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

[†]Ibid, ibid.

[‡]Cf. pp. ·38ff and §1e (pp. ·89 ff) of ch. 3b, and ch. 4.

*«ref»

**deKleer (1986).

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 Artificial Intelligence's original interest in self, cited for example in Collins (1975), are not limited to knowing *what* one believes, but having some *understanding* of it. The potential subject matter of introspection, in other words, should be understand to be at least as broad as necessary to include clinical psychology and psychiatry, and perhaps sociology as well. 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?"

4.b.i 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(α) should be represented just in case α is, and ¬B(α) just in case A38

α is not. This simplicity depends partly on the simplicity of the introspective representational language, but also on another property of S I have not 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 does not 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 cannot be true, because it conflicts with something else I believe" could be straightforwardly 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 cannot 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 $A_i\varphi \supset A_iA_i\varphi$ 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. A39 A40

Truth is not the only significant property, and therefore is not the only aspect of integrity that matters, as we can see by looking at Bowen and Kowalski's DEMO predicate. According A41 A42

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., ⍰DEMO⍰) = ⍰, to be a little cavalier about notation). But this is not all that is required, if DEMO is to play the role that Bowen and Kowalski 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' being (actively) derived from PROG. Similarly, in 3Lisp, where 'φ' was used in the external theory to signify [declarative] content (i.e., roughly ⍰…⍰), and 'ψ' to signify procedural consequence (roughly, ⍰), and where the internal (impression) designing procedural consequence was called NORMALISE, it was necessary to show not only that A43 φ(NORMALISE)=ψ, but also, very roughly (ignoring some use/mention issues) that ψ(NORMALISE)≈ψ. 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 comes to is something like the following. In order to count as having introspective access to some aspect of your self, not only must you be able to *represent* that aspect; you must also be able to *use* that representation—to 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 cannot simulate in virtue of reasoning with those thoughts. But one of the advertised powers of introspection is its ability to enable us to do things differently from how our underlying architecture would have done them [had we not introspected]. If we cannot do them (introspectively) in the same way [modulo timing] that the architecture would have done them (non-introspectively), there seems little chance that we will ever 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 cannot ride a bicycle by thinking about it, my bicycle-riding thoughts do not qualify for the label *causally-connected introspection*.

4.b.ii 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 in the literature 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; 3Lisp's □ and □; etc. In the discussion so far, I have characterized 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

9. 'Linking rule' is used in Bowen & Kowalski (1982), 'semantic attachment' in Weyhrauch (1980), 'level-shifting' in des Rivi6res and Smith (1984) [ch. 5], and 'reflection principles' in Weyhrauch (1980) and some of the meta-logical tradition.

self to representation (I will call this "upwards"), and from representation to self ("downwards")—take on rather different properties. The differences are at least analogous 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 (the wrong thing to do), 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 is no good to think "Ah, yes, the second millennium is drawing to a close," as it was when you fell in; you want to represent the very local situation that led you to fall into the river, represented in the appropriate way. This is the connection from reality (i.e., self) to representation. But similarly, after analyzing the affair, and concluding that things would have gone better if you had leaned the other way, you do not want merely to sit on the bank, fatuously contemplating an improved self: the idea is to get back in the water and do better. That is, you need a connection from representation to reality (more like the situation when you have a goal or even intention): you have a representation, and you want the facts to fit it. Both kinds of connection are germane even for as simple a self-referential representation as $\neg \mathrm{B}(\alpha)$; the system might need to know whether $\neg \mathrm{B}(\alpha)$ is true, or it might want to make it true. On S's reading of B as "is explicitly represented" neither direction 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 is not computable, even upwards. In the downwards case, moreover, if the prop-

†Included here as ch. 4.

erty represented is a relational one, there may be no unique determinate solution (lots of things, typically, could make ¬M(α) 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.

Since this problem is simply a particular case of the general issue of designing and planning action, however, and not specific to the introspective case, it need not concern us more here.

4.b.iii 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-EXPLICITLY-REPRESENTED-BELIEFS(6)

Instead, one would need to add

TOTAL-NUMBER-OF-EXPLICITLY-REPRESENTED-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: overall, the implicit dimension or aspect of the system that was rendered explicit remained the implicit dimension or aspect of the explicit rendering. There was no circularity involved, but there was overlap, with concomitant problems.

Overlap is not 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* or *layered set of implicit contexts*, analogous to that provided by type hierarchies in logic, so that the introspective process can muck about with its subject matter without affecting the circumstances that give that subject matter its content.

Overlap only arises when the introspective machinery makes explicit some implicit aspect of the internal circumstances; it is not a problem when what is implicit to the base-level is also implicit for the introspective machinery. Thus vari-

ous systems, such as MRS and Soar, apparently do not make explicit any otherwise implicit state (everything that can be A44 seen, self-referentially, is *already* explicit; what is implicit remains so), so the problem of overlap does not arise. In some other cases, such as in BROWN,† 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 3Lisp 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.‡ The continuation structures of 3Lisp, 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—did not overlap (this is why, as we said earlier, an introspective variant of 3Lisp that only rendered these two aspects explicit would be essentially trivial).

3Lisp'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 3Lisp—because, as stated, continuations are not 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

†Friedman and Wand (1984).

‡At the time of its design I called 3Lisp *reflective*, not *introspective*, but I now think this was [at least partially mistaken]. Reflection—see below—was what I wanted, but introspection was what I had. A45

particular architecture, however, will clearly not 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, although 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 is not exactly easy to come by accurate 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 theorized about, but from the difficulty of obtaining a non-indexical representation of one's participation in the external world.

4c. Reflection

In the last section a point was made that I 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 the single-argument 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 A46
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 will not 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 is not universally shared [or anyway shared by one's peers]. 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 will not realize. Introspection is therefore a prerequisite for substantial reflection (self-knowledge is a precursor of detachment, as history has repeatedly told us). 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 have been tracking, neither the autoepistemic nor the control has addressed relativity to the external world at all. 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 has been closely based on, mathematical logic—which, as Barwise and Perry have repeatedly emphasized, does not admit, in its foundations, of external relativity to circum-

stance. 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 A47
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 external 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, behavior. Although one
can (and I do) argue that the resulting computational systems A48
are themselves representational, and therefore bear a "content" relation to the world in which they are ultimately deployed, that system-world relation is not addressed by traditional programming language analyses. As a result, the implicitness represented by such self-referential models as meta-circular interpreters, BROWN, MRS, etc.,† is also primarily internal.[10]

†Steele & Sussman (1978), Friedman and Wand (1984), and Genesereth et al. (1983), respectively.

10. Not realizing this fully at the time, I did not initially describe 3Lisp (Smith 1982, 1984) [chs. 3 and 4] in a way that was very accessible to the programming language community. 3Lisp'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 behavior was being engendered, but also the subject matter of the resulting system. I still believe that this is often how programming is understood,

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, to the extent that it is causally effective at all, 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 is not 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.

None of this means that serious reflection is impossible, however, partly because of our three-way, rather than two-way, categorisation of circumstance into external, indexical, and internal types. The truth of whether Shakespeare wrote the sonnet is external; the implicitness motivated by efficiency, in contrast, is typically indexical, not external, and indexicality has to do with the circumstances in which the agent participates—which circumstances, some of which, at least, should be relatively *nearby*. If there is any locality in this world, there

even if implicitly, by a large number of programmers: my analysis; however it 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 3Lisp was able to deal with was that of pure (and simple) mathematics, so it did not really live up to its own semantical mandate.

seems more hope of an agent's knowing about local circumstances than about situations arbitrarily remote in space and time. What is 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: (i) because of the complexity of the calculations involved, such as those illustrated by consistency; (ii) because of the theory-relativity—no theory can render everything explicit; and (iii) because some circumstantial relativity—particularly indexical and external—remains beyond the causal reach of the agent. But there are other limits as well, An important one stems from the fact that the self being represented is ultimately the same self as the one doing the representing, and as such certain possibilities are physically [if not metaphysically] 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 in climbing up out of its embedded position, a system 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. A49

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 one-

self. 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 is 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', recognizes its special self-referential role, collapses back down to the fully implicit structures that engender talking, cuts them off, and thereby shuts up.

That is 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 traveler in the upper reaches of 3Lisp, 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, has benefited from his works. My thanks.

A50