THEORY AND APPLICATION OF SELF-REFERENCE: LOGIC AND BEYOND

TO APPEAR AS CHAPTER IN CSLI BOOK.

THEORY AND APPLICATION OF SELF-REFERENCE: LOGIC AND BEYOND

DON PERLIS

ABSTRACT. This paper begins with a brief (and probably rather personal and one-sided) overview of self-reference as seen in logic and AI, then slides over to speculations on self-reference in natural language, and eventually returns to AI and even more speculative theories of the conscious mind based on one kind of self-reference (arguably the most fundamental kind).

1. INTRODUCTION

Self-reference has been a topic of study in formal philosophy and logic for a great many years, indeed for millenia. Yet, I shall argue, many ostensible cases of self-reference so studied are not, in fact, cases of reference at all. To get to reference, and to self-reference, one must go beyond the usual cases, formal and otherwise. To present this argument in context, I will begin with a brief review of some of the more familiar examples of (so-called) self-reference and the associated puzzles and problems ("conundra") that have drawn attention to them. I will then discuss a particular solution to these conundra, based on work of Gilmore and of Kripke, a solution that seems to avail itself of a "normal order principle". This in turn will lead to the question as to how satisfactory such a principle is, in light of further examples from natural language and from cognitive philosophy. Finally, some tentative conclusions are drawn, about the nature of language, reference, and self.

2. SOME CONUNDRA

Perhaps the most famous example of a self-referential utterance is the so-called *Liar*:

This sentence is false.

The Liar, or L for short, has two curious features: (i) it appears to refer to itself, and (ii) it appears to contradict itself. The latter feature is the one that has received the vast bulk of attention, and we shall give that aspect attention as well.

However, we also will critically examine the former feature, in stages, paying special attention to the word "This". At this early stage let us simply note that there is an issue as to what, if anything, can guarantee that "This" in L succeeds in referring to L itself, as opposed (say) to some other sentence that may have recently been uttered or pointed to. An altered "display" version is

$$DL: \qquad\qquad DL \text{ is } false.$$

But what guarantees that the name "DL" in the display refers to the sentence

that follows it? Presumably it is our agreement that it so refer; but then there is personal agency involved.

Perhaps "This sentence that I am now writing/uttering/expressing, beginning with the word *This*, is false" will provide a more convincing guarantee, wearing, so to speak, its meaning on its sleeve. But this then rests upon a clear reference for "I", presumably once again an agent with referential *intentions*, a matter seemingly far removed from the usual concerns in discussions of the Liar. Another sort of guarantee might be along the lines of "The only sentence written on the whiteboard at 4:19pm on May 12, 2003, in room 3259 of the A. V. Williams Building at the University of Maryland, is false." We will set these new concerns aside for now, returning to them nearer the end of our essay.

The usual concern with L is that it appears to contradict itself: If L is true, then what L says is true, i.e., L is false, so L cannot be true. But if L is not true, then L is false, and this is what L says, so L is true after all. Thus it appears that L can have no consistent truth-status. Yet then, in particular, L is not true, and so what it says again seems true.

Perhaps "not true" and "false" should be distinguished. Yet "not true" is what "false" means, by common accounts. And in any case, we can recast the Liar as sentence L':

This sentence is not true.

Now the distinction between "false" and "not true" is irrelevant, and yet the conundrum remains.

One natural-enough reaction to the above circumstances is to suppose that some sentences are, after all, neither true nor false, indeed neither true nor not true. This may seem to play fast and loose with the word "not"—after all, "not" simply asserts the failure of what follows it. But perhaps there is something hidden here. In order to be a candidate for failure in the matter of its truth, and sentence must first have a potential truth that can fail, i.e., it must have a clear enough meaning that can be measured against some criterion of truth. How does a sentence acquire a meaning? This is the subject of much dispute, and we will return to it a bit later on. But what we need now seems more modest: the notion that meaning—whatever it is—allows a distinction between sentences S for which there is a clear separation between the meanings of

S is true.

and

S is not true.

and those for which there is not such a distinction.

Not all sentences meet this (admittedly imprecise) "separation criterion". The following examples may help clarify the point:

Colorless green ideas sleep furiously.

That sentence is false.

This sentence no verb.

The first of these is a famous example due to Chomsky; it illustrates a grammatically (syntactically) correct English sentence, but one that has little if any meaning (semantic nonsense): one would be hard-pressed to say what it means, and therefore whether it is true or not. The second example seems unproblematically meaningful,

yet without further information as to what "That" refers to, it cannot be said to be either true or false. The third example is not a sentence, strictly speaking, but it has an emphatically obvious meaning, and indeed one that is true.

The present point is that, as in the second example above, a sentence may be neither true nor not true, in the sense that its meaning does not make a clear enough separation between these. In the case of the second example, the lack of separation appears to rest on missing external information (what "That" refers to). The Liar, on the other hand (assuming we take its "This" as unproblematically self-referential) fails to make a clear enough separation due to an internal circularity. In that respect the Liar is much like the $TruthTeller$:

This sentence is true.

which (although not contradictory) also appears not to make a clear enough claim to determine it to be true or not: its truth seems to presuppose its truth, so we never manage to pin it down or separate it from its failure.

This suggests a *Normal Order Principle*: Truth (of a sentence S) is a relation between the world, the sentence S, and the meaning m of S, where the relation has a temporal character: W precedes S which in turn must precede m (a relation between S and the world), which in turn precedes the truth (or lack thereof) of S. This—also sometimes referred to as *grounding*—will be the subject of a later section. Here we shall instead briefly mention a few other matters, formal and otherwise.

The examples we have considered so far are informal, based largely on common-sense notions. However, it is not hard to capture similar behaviors in more formal dress. A key component of the formalization is the Diagonal Lemma, which asserts that in any reasonably expressive formal theory F, for each unary wff Px there is a sentence p such that it is provable in F that $p \iff \neg P'p'$. Here $'p'$ is a name for p, i.e., a term that allows us to predicate P of p. Given the Diagonal Lemma various results follow, some more easily than others:

(1) *Schema T* (which says that, for all wffs, a $T'a' \iff a$) is contradictory. Here T is to be thought of as a truth-predicate, asserting the truth of its argument. But applying the Diagonal Lemma, we get, for a suitable wff t, $t \iff \neg T't'$ and so by Schema T, $t \iff \neg t$, a contradiction.

(2) From the above result, renaming t as L, we have

$$L \iff \neg T'L'$$

a formal counterpart of the Liar: L is equivalent to its own denial. And as just seen, this leads to a formal inconsistency.

(3) With more work (Gödel) one can devise a wff Thm such that, for all sentences a, $Thm'a'$ is provable in theory F if and only if a is provable in F; that is, Thm behaves like a provability-predicate. But by the Diagonal Lemma we also have that, for some wff g, $g \iff \neg Thm'g'$; i.e., g is equivalent to its own unprovability in F. It follows that if g is provable, so is $\neg g$. Hence we get Gödel's Theorem: g is either unprovable (and hence true, in the sense that what it "asserts"—its own unprovability—holds), or F is inconsistent.

Another example brings us closer to artificial intelligence, where what a reasoning agent can be said to know is of great interest. Suppose there to be an very wise

agent, A, who has excellent powers of self observation, including knowledge of what A does and does not know. That is, A knows a fact f if and only if A also knows that A knows f, and moreover, if A does not know f then A knows that A does not know f. Formally (but suppressing quote marks for readability):

$$K_A(f) \iff K_A(K_A(f))$$

and

$$\neg K_A(f) \iff K_A(\neg K_A(f))$$

Suppose we also regard A's knowledge as being what A can *prove*, so that, in effect, A is a formal theory. Then K behaves like a strong version of Thm, i.e., the above schemata become

$$A \vdash f \iff A \vdash K(f)$$

and

$$A \nvdash f \iff A \vdash \neg K(f))$$

But now we can show that A is inconsistent.

From the Diagonal Lemma we get a wff k such that $\vdash k \iff \vdash \neg K(k)$. But from the above schemata, if k is not provable, then $\vdash \neg K(k)$ and then $\vdash k$, i.e., k is provable after all; so it cannot be that k is unprovable. But then k is provable, and so $\vdash K(k)$, and also $\neg K(k)$; thus A is inconsistent. This we may call the "No perfect-self-knowledge theorem". It indicates that artificial agents (as well as "real" agents) are subject to certain formal constraints on their knowledge.

All the above *formal* results, it should be admitted, are purely syntactic, and reference (let alone self-reference) plays no real role. For example, that L refers to anything is irrelevant to the above proof that the formal *Liar* is inconsistent. And the Gödel sentence g does not really refer to anything at all, let alone to its own unprovability.

And who cares? Isn't self-reference just a curiosity?—an accident allowed by a side-effect of over-expressive language, with surprisingly useful but equally accidental application in formal logic, and of no deep significance in itself? Cannot we then simply ignore the bad (contradictory) cases (e.g., Schema T which allows the *Liar* in full force) and welcome others (such as provided by Gödel)? Perhaps, but we shall argue otherwise. First however we shall look at some standard methods for isolating the bad cases: the Tarski Hierarchy, and the Gilmore-Kripke partial models that appear to provide a formal characterization of Normal Order.

3. The Tarski Hierarchy

Since the above formal results are just that, formal (syntactic) and not dependent on semantics, then it might be possible to keep the advantages of certain "seeming" self-reference (as in Gödel's powerful Theorem) without the disadvantages of the *Liar* (such as inconsistency). Tarski showed how to do this by means of a restriction on how languages refer. He posited a hierarchy of languages $L_1, L_2, \ldots$ where each L_{j+1} has expressions that refer only to objects in a previously defined language L_j. There is then no expression that can refer to itself, and no truth predicate T that can be applied to an expression that also has T in it. Rather, each L_j has its own truth-predicate T_j that can be applied (only) to expressions in L_{j-1}. This approach simply banishes self-reference from expression altogether, while leaving intact the

syntactic vestiges needed for the Diagonal Lemma (and useful formal results such as Gödel's Theorem).

Thus in particular, Schema T is banished, and with it the possibility of the *Liar* and contradiction. On the other hand, the Hierarchy seems to banish too much. There are perfectly innocuous but semantically-based cases of self-reference, such as:

This sentence has five words.

Yet this is not expressible in the Tarski Hierarchy. Nor is the following pair of straightforward sentences, each referring to the other (one happens to be false):

The sentence below has seven words.

The sentence above has six words.

Note that the Hierarchy banishes statements that have a sentence S appearing both as S *simpliciter* and as a named object $'S'$ together, for then the latter would refer to the former, a violation of the hierarchical order. But the Normal Order Principle mentioned earlier does allow *expression* of such statements; it simply clamps down on how a statement with a truth-predicate can itself be judged true or false: namely, when there is sufficiently clear separation between these and, further, when the truth-judgment is made *after* the meaning is determined.

4. The Gilmore-Kripke Approach

Let us review the Normal Order Principle. A paradigmatic case is this: *Snow is white.* Here there is a fact of the matter (that crystalline water tends to reflect fairly uniformly across the spectrum of visible light) and there is also a sentence (the one just cited); and there is a connection between them that allows (in principle) its verification once the former ingredients are in place. We have the world, then we have the sentence, and then we ascertain a potential connection between them that distinguishes two cases (true, false), and finally we judge one of these to obtain. If instead one sentence refers to another (rather than, say, to snow), then the referring sentence treats the referent sentence as part of the world. Put yet another way: we look to see (e.g., that x), then record the result $True(x)$. The temporal order is this: first there is (i) the world W, next (ii) a sentence S, then (iii) a connection of meaning between S and W, and finally (iv) a determination of truth or falsity with respect to that meaning. If this process fails, then we have nevertheless learned something about S, namely that it is neither true nor false—which is simply to say again that the process has failed!

In other words, the Normal Order Principle defines truth and falsity as successful outcomes of this process. As such it seems natural enough, at least at first glance. And it has a formal counterpart in work of Gilmore [4] and Kripke [7], who (independently and differently) provided analyses of certain conundra and tools for addressing them, along such lines. Gilmore dealt with set-theoretic versions of the *Liar*, such as *Russell's Paradox*: is the set R whose members are all those sets that do not contain themselves as members, a member of itself? His treatment focuses on the set-membership relation $x \in y$ and a special method for constructing models in partial steps, akin to the Normal Order Principle. This leads to a consistent formulation in which nevertheless the Russell set can be defined, but the question as to whether it is or is not a member of itself is, in effect, not well-defined. That is, $R \in R$ does not obey Normal Order (for set membership rather than truth).

This does not mean that no such set as R exists; on the contrary, in Gilmore's set theory it is provable that R exists. Kripke provided a similar treatment for truth. Later Feferman [2] and Perlis [10, 11] also working independently, unified these two treatments. The following schemata capture much of this approach:

$$T(\alpha) \iff \alpha*$$

where $\alpha*$ has no effect on α unless there is an embedded and negated T inside, and then:

$$(\neg T\beta)* \iff T\neg\beta$$

Finally we also require

$$\alpha* \implies \alpha$$

That is, the assertion of the truth of the negated truth of an embedded statement, amounts to the assertion of the truth of the negation of the embedded statement. Put differently, T pushes negations through embedded T's. This seemingly convoluted affair, in "everyday" cases of β, reduces to a triviality:

$$T(\neg T\ 0 = 1) \iff T\ 0 \neq 1 \iff 0 \neq 1$$

Thus—reading from right to left above—the bare fact of $0 \neq 1$ then gives rise, in normal order, to the assertion that it is *true*, and so on. But for certain other cases things are more interesting. For instance, here is what happens in the case of the *Liar*:

$$L \iff \neg T(L)$$
$$T(L) \implies T\neg TL \implies T\neg L \implies \neg L$$

yet also $T(L) \implies L$; so T(L) is contradictory and thus $\neg T(L)$ is proven. Also, $T\neg L \implies \neg L \implies TL \implies L$ so $T\neg L$ is contradictory and thus $\neg T\neg L$ is also proven. Thus neither L nor $\neg L$ is true, even though L is provable (it is equivalent to $\neg TL$ which we just proved). Thus a provable wff need not be true (in the sense of the T-predicate). Curious, but not contradictory. The attempt to unpack L into a bare fact that stands independently prior to assessing $T(L)$ leads nowhere, or rather, leads in a circle in such a way that we can actually prove it cannot get a positive truth-judgment.

Of course we would expect the above (from the Normal Order Principle), since neither L nor $\neg L$ separates enough to allow its truth clearly to obtain or clearly to fail. And the underlying formal treatment, following Gilmore, is provably consistent, so it will not happen later on that someone will discover a new conundrum (formalizable in the Gilmore-Kripke setting) that leads to an outright inconsistency. Thus one can have the *Liar* and pacify its tendency toward contradiction too. The above two schemata for the Gilmore-Kripke approach (GK for short) form a modification of Schema T, one that appears to capture a sense of T that comes close to the intuition behind the Normal Order Principle.

5. Flaws

The $TruthTeller$ ought similarly to allow a proof that it does not come out true, on the intuition underlying GK, that $T(x)$ signifies that x is judged true *after* x is ascertained. Since one cannot ascertain $TruthTeller$ before it is judged true (these are one and the same thing) then the normal-order test fails, and so $\neg T(TruthTeller)$ ought to result (and with it, $\neg TruthTeller$). But formal means

to show this failure of normal-order-ascertainment for this sentence, are not present in existing treatments along the lines of GK.

Recall that the Tarski Hierarchy provides no means to express

$$\textit{This sentence has five words.}$$

It is not that there is a difficulty in counting five words, nor of recognizing (parsing) a sentence. "Five" and "sentence" can be given adequate formal "meaning" by appropriate axioms allowing derivation of the desired results. The difficulty is in associating *that very sentence itself* with the phrase "This sentence". We can finesse it *a la* Gödel, using a special name or number $'S'$ for a particular sentence S. That is, we can form a predicate $Five$ such that $Five(x)$ means x is a sentence with five words, and via the Diagonal Lemma there will be a sentence S such that $\vdash S \iff Five(S)$. But the embedded S (in suppressed quotes) is a problem; it inhabits a level that must come before the level of the main sentence, and yet at the same time it occurs at that later level as well, a clear violation of the Tarski Hierarchy.

Again, the underlying intuition behind GK presumably should provide a solution to this problem by allowing a sentence S such as above—in the version in which $\vdash S \iff Five(S)$—to exist as an object before its meaning (or truth, or fiveness) is determined. Then S can be counted to see whether $Five(S)$ holds. This even though normal order has not entirely been respected here: we look to see, then we should record the result, but the result is already recorded before we look. Still, the *meaning* of the $Five$ sentence can be assessed, and its truth judged, after the sentence has been recorded. The attempt to assess and judge does not circle back to that very attempt, dooming it to failure; rather it leads to a counting and a definite unique conclusion, even if that conclusion happens to be already written down and indeed happens to be the very object being counted. Yet no current GK-style formalization appears to have the requisite machinery to capture this phenomenon.

The situation so far is this: GK allows standard examples of seeming self-reference to be "pseudo-"expressed, without actual inconsistency, and in rough accordance with an intuitively sensible principle. But there are flaws. The first is simply that certain examples (such as the $TruthTeller$ and the $Five$ sentences) fit the intuition behind the principle but the formalization does not fully oblige, as we have seen.

The second flaw is that the very pseudo-expressings that allow GK to (appear to) treat self-reference without banishing it, leave out the ingredient of (genuine) reference. Finding a sentence S that is provably equivalent to, say, $P(S)$, does not in itself indicate that S refers to anything at all, let alone to itself. The deictic "this" of natural language (as in "This sentence is false.") has been by-passed altogether in formal treatments. Indeed, reference (or semantics) of any kind is traditionally placed *outside* a formal language, as a function defined on expressions in the language, mapping to an external domain. The language in question does not typically have an expression that stands in for this function; and even if it did, what would determine that standing-in relation? It is as if meaning, or truth, is always one step removed, leaning on some agreement lying outside whatever language is used. The GK approach does have a truth-predicate, but it relies on an external naming convention that leaves meaning, as a link between expressions E and referents O, unexpressed.

Regarding the first flaw, perhaps a remedy can be given in an extended formalization of GK. Perhaps introduction of a $Means$ predicate along with some representation of temporal ordering can provide a more explicit formal representation of (a version of) the Normal Order Principle, allowing a treatment of the $TruthTeller$, as well as the $Five$ sentence, to follow intuition.

In regard to the second flaw, it is only by means of an agreement among whichever logicians happen to be participating in the discussion, that $'S'$ refers to anything at all, let alone to the sentence S, given, say, that $S \iff T(S)$. For the $\iff$ relation is not the same as a referring relation (to self or anything else). After all, $0 = 1 \iff$ *Snow is composed of carbon*, but $0 = 1$ does not thereby refer to anything, least of all to snow and carbon. Perhaps, again, some of what is missing can be restored by introduction of a $Means$ predicate, as in $Means(E, ...O...)$ to indicate (among other things) that expression E refers to object O. (But even this would beg the question as to who or what takes $Means$ to so indicate—indicating already being a kind of referring or meaning).

It would be of interest to formalize such a notion of meaning, as an underpinning on which to analyze truth as a derivative notion. For example, one might attempt to characterize $T(x)$ in terms of the meaning $m = (m_T, m_F)$ of x where m_T and m_F are sets of possible worlds (where x is "true" or "false", resp.), W is the real world, and $W \in m_T$. Then x might have a meaning even if it is not separate enough from that meaning to be judged true or false (that is, if W is in neither m_T nor m_F). Thus one would anticipate that for $tt = TruthTeller$, W would come out in the middle (tt being neither true nor false in *any* world)—as it already does in standard GK—and moreover that this fact would be noted as $\neg T(tt)$ and $\neg F(tt)$, much like the $Liar$.

But meaning is a bit more tricky than this. For instance, $Blue(a) \wedge \neg Blue(a)$ certainly seems meaningful, indeed quite definitely false. And its meaning seems just as definitely different from that of, say, $1 \neq 1$, which is also false. Thus neither sentence has a model in the conventional sense, and so neither would be true in any world, and both would be false in all. Their meanings (in terms of pairs m_T and m_F) then would be identical. This is a complex issue (see [3, 8]); we will here only hint at an idea that might be worth further exploration: Perhaps a semantics can be developed that avails itself of superposed worlds, to allow one to conceptualize a meaning for expressions such as $B \wedge \neg B$, e.g., a pair of worlds where B holds in one, and $\neg B$ in the other. Such a pair for $Blue(a) \wedge \neg Blue(a)$ would not generally be the same as the corresponding pair for, say, $1 \neq 1$. But the pair for the latter presumably would be the same as for $1 = 1 \wedge 1 \neq 1$, namely $m_T =$ all worlds, and $m_F = \emptyset$, which seems satisfying.

Thus there is hope that the first flaw can be repaired, and the second at least partially addressed, by taking meaning more seriously as a concept to be formalized (and semanticized). But the deeper aspect of the second flaw remains: any kind of map between symbols and referents is arbitrary, leaning on a decision to use that map, and not some other, but an agent who intends to use that map. Only when this issue is faced head on, do we encounter genuine cases of reference, and the possibility of genuinely self-referring expressions.

6. No representation without representers!

A lesson we draw with respect to the second flaw above is this: self-reference proper has largely been left untouched by the very large literature purportedly on that subject. This is because *reference* has largely been left untouched, or rather pushed to the sidelines, via Gödel numbering or a similar artificial mechanism that leans on external agreements to bring reference in at all. Attention has focussed, rather, on formal counterparts of self-reference that do, to be sure, carry with them a substantial potential for contradictoriness, in close analogy to their informal—but more genuinely self-referential—sources. But while this attention has produced much of great importance, it has left much out as well. First and foremost is this problem: can there be representation (meaning, reference) without an agent who chooses to so represent? And secondarily, what is the relation between reference in general, and self-reference? Third, what can be said about "genuine" informal self-referential expressions, in light of answers to the former questions?

We will not attempt here to give a detailed analysis supporting a view with respect to the first two questions. We will however state a position (referring the reader to [13, 14, 15, 16] for such an analysis): Reference of E to O depends inherently on a referring agent A to make the connection between E and O. That is, the referring of E to O is performed *by* A rather than by the expression, E. Moreover, to perform a referring act, the agent A must *take itself* to be so referring, as part of that same act. Thus a referring act is self-referring, and so self-reference is a necessary component of reference of any kind. This position is, admittedly, not coin of the realm. But I believe it to be the only way to address head-on the underlying issues. Some further motivation will be found by a brief return to some earlier examples:

N: This sentence that I am now writing/uttering/expressing, beginning with the word "This", is false.

W: The only sentence written on the whiteboard at 4:19pm on May 12, 2003, in room 3259 of the A. V. Williams Building at the University of Maryland, is false.

The seeming contradictoriness of N (for Now)—and also of W (for $Whiteboard$) if that happened to be the only sentence on that whiteboard at that time—is not the issue here. The issue rather is what makes these sentences self-refer. In the case of N it appears straightforwardly to be the presence of a self-referring agent "I". In the case of W there is no immediately apparent agent, but without an agent (or even a whole linguistic community of agents) *somewhere* in the broader setting, these marks on a whiteboard have no meaning at all (after all, they are in *English*). And that agent (or agents) will have to self-refer in order to make sense of either N or W, e.g., to identify the University of Maryland as located in real physical space situated somehow with respect to their own locations.

7. A highly non-normal order

We are now moving from the speculative to the "more speculative", away from formal logic and toward philosophy.

A genuinely self-referring expression (or self-referring act) would appear to severely violate normal order: it would refer to itself at the very same time as it is expressed (or enacted). That is, its very expression (or enactment) would amount

to its self-reference, making a temporal sequencing seemingly out of the question. Here are two more examples:

I am now, with this very utterance, speaking English.

This is a hat.

The latter does not seem to fit, unless we regard it as a gloss for "This object I am calling your attention to now, with this very utterance, is a hat". Such a view is—or is close to—one taken by Grice [5], to the effect that all utterances surreptitiously self-refer. Related observations have been made by Millikan [9] and Perry [17]. On the basis of such considerations, it is argued in [13, 14, 15] that *all* intentional utterances are self-referential via their self-referring utterer. That is, all cases of genuine referring are cases of an agent who intentionally uses an expression to refer, and who in so doing performs an act of self-referring akin to the hat example above: "the idea that I am intending, in this very act of expression, is such and such."

The above brief description of this claim does not do justice to the underlying idea, and the reader is referred to just-cited papers for additional argument. There it is further hypothesized that thought in general, and not just communicative utterances, carry (and do so more deeply) the underlying weight of the self-reference. To think a thought *is* to think its meaning, i.e., thoughts carry their own meanings (whence the violation of normal order). One can then formulate various simple examples of self-referential *thoughts* vaguely analogous to the *Liar* and the $TruthTeller$, e.g.:

This is a thought.

This is not a thought.

I am (right now) thinking.

I am not (right now) thinking.

Presumably—if taken as expressions of actual ongoing thoughts in an agent—the first and third of these are necessarily true, and the other two necessarily false. We can speculate further: as long as we are aware of anything at all, to that extent we have a thought (to wit: there is an X), and then there is an implicit "That thing (that I am hereby noting before me) is an X" and so on. A self-referring self seems to underlie all thought or awareness. Thus a bare-bones agent self-reference may be the most basic kind of reference. What is bare-bones self-reference like? Imagine yourself stripped little by little of this sensation, that thought, until all that is left is your own grasp of being awake but not aware of anything else. This may be what is suggested in Piet Hein's poem "EVENING AND MORNING SONG—About falling asleep and waking up" [6], with a dimming of awareness upon falling asleep; or with the first glimmer of awareness upon waking up, in which one has not yet recalled one is the person of yesterday with plans for the day ahead, not yet at first identified with a name or a history or a persona beyond the primitive self comprised only in the self-referential and self-creating thought that *this self-awareness is*:[1]

> The world disappears,
> a loop running smaller, until
> the thread is drawn out,

[1]My thanks to Torkil Heiede for bringing this poem to my attention.

and the space it encloses is nil.

Newborn of nothing,
reluctantly starting to be,
fumbling awareness awakens
and finds that it's me.

Such ideas, spun out far enough, have led to the hypothesis [15] that agent self-reference not only underlies all reference, but indeed is tantamount to conscious awareness. Let us step back a little from such tenuous speculations, to a more engineering perspective, pursuing an idea of John Perry [17]. There Perry describes pushing his shopping cart along in an attempt to find the shopper whose cart is leaving a trail of sugar on the floor, only later to realize *he* is that shopper. Setting aside various philosophical issues here, we end our discussion by posing related questions about robot design. Consider a robot that can decide *it* is the robot who is, say, leaking oil, upon hearing that robot #17 is leaking oil. What is it for robot #17 to know that it, itself, is that robot? How does this affect its behavior? Presumably it is quite important to have such a capability, e.g., for survival. See [12, 1] for more elaborate discussions of this idea.

In conclusion, the topic of self-reference appears to span a vast intellectual territory, from formal logic to natural language, to philosophy of mind, to artificial intelligence and robotics. And very many open questions remain.

References

[1] Michael Anderson and Don Perlis. The roots of self-awareness. Forthcoming.

[2] S. Feferman. Toward useful type-free theories–I. *Journal of Symbolic Logic*, 49:75–111, 1984.

[3] G. Frege. Uber sinn und bedeutung ('On sense and denotation'). *Zeitschrift für Philosophie und philosophische Kritik*, pages 25–50, 1892.

[4] P. Gilmore. The consistency of partial set theory without extensionality. In T. Jech, editor, *Axiomatic Set Theory*, pages 147–153. Amer.Math. Soc., 1974.

[5] H. P. Grice. Meaning. *Philosophical Review*, 66:377–88, 1957.

[6] Piet Hein. *Grooks IV*. Doubleday, New York, 1972.

[7] S. Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975.

[8] Alexius Meinong. *Uber Möglichkeit und Wahrscheinlichkeit*. Leipzig: Barth, 1915.

[9] Ruth Garrett Millikan. *Pushmi-pullyu Representations*. Ridgeview Publishing, 1996. Chapter in *Philosophical Perspectives vol. IX*, ed. James Tomberlin.

[10] D. Perlis. Languages with self reference I: Foundations. *Artificial Intelligence*, 25:301–322, 1985. Reprinted as a chapter in *Reflexivity: A Source-Book in Self-Reference*, S. J. Bartlett (ed.), North-Holland, 1992.

[11] D. Perlis. Languages with self reference II: Knowledge, belief, and modality. *Artificial Intelligence*, 34:179–212, 1988. Reprinted as a chapter in *Reflexivity: A Source-Book in Self-Reference*, S. J. Bartlett (ed.), North-Holland, 1992.

[12] D. Perlis. Intentionality and defaults. *International J. of Expert Systems*, 3:345–354, 1990. Special issue on the Frame Problem, K. Ford and P. Hayes (eds). Reprinted as a chapter in *Advances in Human and Machine Cognition, vol. 1: the Frame Problem in Artificial Intelligence*, K. Ford and P. Hayes (eds.), JAI Press, 1991.

[13] D. Perlis. Putting one's foot in one's head—part I: Why. *Noûs*, 25:435–455, 1991. Special issue on Artificial Intelligence and Cognitive Science.

[14] D. Perlis. Putting one's foot in one's head – part II: How. In Eric Dietrich, editor, *From Thinking Machines to Virtual Persons: Essays on the intentionality of computers*. Academic Press, 1994.

[15] Don Perlis. Consciousness as self-function. *Journal of Consciousness Studies*, 1997.

[16] Don Perlis. What does it take to refer? *Journal of Consciousness Studies*, 2000.

[17] J. Perry. The problem of the essential indexical. *Nous*, 13:3–21, 1979.