Annotations[1]

A1 ·1/1/7:9 The claim that the "relativity of representation to circumstance facilitates local inference, and enables representation to connect with action" is metaphysical: it is because of the fundamental deixis of the underlying (physical) plenum. Deixis is not a property I gave explicit voice to until 1996, in o3—though there is some discussion here in §2c ("Efficiency") on pp. ·12–·15. See also "Rehabilitating Representation," ch. ... of Volume II.

A2 ·2/1/4:5 Re "naming himself as part of his subject matter": I mean, very simply, his use of the first-person pronoun 'I'.

A3 ·4/1/5:6 As is evident from the discussion in the rest of this paragraph, the claim that "[t]echnical results in the three traditions overlap surprisingly little" was made with reference to investigations of self-reference in Artificial Intelligence, but even now (in 2012) the point remains largely true—not only of relations between and among disciplines, but to a considerable extent within them as well. Even in philosophy, for example, studies of the logical paradoxes of self-referential sentences,[2] considerations of the relation between higher-order and quotational forms,[3] analyses of ramified type theories,[4] explorations of non-well-founded set theory,[5] discussions of personal identity and self-knowledge,[6] phenomenological treatises on Being, etc., are not only likely to be formulated in separate vocabularies, but also to be conducted within relatively distinct philosophical subcommunities. The situation is even more dramatic when one crosses disciplinary boundaries—e.g., to include studies of reflexivity in science and technology studies,[7] varieties of ego, self, and identity explored in psychoanalytic theory,[8] self-referentiality and reflexivity in 20th century literature and art,[9] and so on. No one could plausi-

1. References are in the form page/paragraph/line; with ranges (of any type) indicated as x:y. For details see the explanation on p.·...
2. «Refs: e.g., to Barwise & Etchemendy's The Liar, Smullyan, etc.»
3. «Refs; e.g., to Kripke's "Is There a Problem with Substitutional Quantification?" [or whatever it was called]»
4. «Refs; cf. Russell originally of course, but ... »
5. «Cf. Peter Aczel, and subsequent literature?»
6. «Refs: John Perry, at least»
7. «Refs»
8. «Refs: check with Diane, maybe»
9. «Refs: check with Stefano?»

bly teach a graduate course on self-reference without first locating it within one or at most a very few of these traditions.[10]

As mentioned in the Introduction, I believe that one of the most important but as yet insufficiently appreciated philosophical consequences of AI (and computation more generally) is the fact that, in virtue of recruting tools to deal with complexity (drawn in large part from computer science), it is able to bear down, in the sort of detail that as philosophers we have become used to in such arenas as logic and set theory, on much more substantial and wide-ranging issues of self, identity, self-knowledge, and the like. The point is manifest from the fact that computational systems could both be illuminated by, and shed illumination on, all of the forms of self-referentiality listed in the previous paragraph.

The difficulty is that, perhaps as a result of need to encompass this very complexity, AI's theoretical understanding of such systems has (at least to date) remained rather shallow—certainly lacking the depth and trenchency that would be required in order to satisfy the traditional normative standards of at least analytic philosophy. At the same time, even if it does so primarily in as yet unarticulated ways, AI also manages to be surprisingly broad, detailed (if not deep), and in some inchoate and unreconstructed ways surprisingly nuanced.

A4 :4/-1::5/0 As noted in the Cover, there are places (such as in this paragraph) where the writing in this paper is woefully amateur. The concerns I was reaching towards, though, still strike me as important—including the issue pointed at here: about how self-reference is not merely internal or self-indulgent, but ties into general "external" capacities for world-directedness (as for example reflected in the title of Perry's "The Essential Indexical" [e.a.]). Part of the motivation for writing this paper stemmed from a commitment to show in detail how technical capacities for for self-reference—ranging from narrow capacities to convert egocentic to allocentric coordinates to such wide-ranging issues as those implicit in the Delphic injunction to "know thyself"—underwrite a general ability to be oriented towards

10. I say "graduate course" only because one could imagine an undergraduate course, of a sort growing in popularity, that takes self-referentiality as a theme cutting across the entire intellectual pantheon, and that uses it as an organizing principle to map intellectual history, or as a framework for a general education or "Great Books"-type of liberal arts foundation course.

"that which transcends the subject," in ways that give rise to consciousness, underlie ethics, etc.[11]

A5 ·5/1/5:8 Re the repudiation of context and circumstance in logic: this paper was written in the midst of my collaboration with Jon Barwise and John Perry on the "situated language" project at the Center for the Study of Language and Information (CSLI) at Stanford. Not just that particular project, but the Center as a whole,[12] grew out of a shared commitment to take context and circumstance seriously. It was our common belief that contextual/circumstantial dependence was a central and constitutive feature of human intentional pratice, rather than, as at the time seemed generally assumed, an undeniable but peripheral complexity.

For a number of reasons, our projects increasingly diverged following the Center's launch. In part, I was far more concerned than Barwise and Perry with issues of architecture and computational efficacy. More profoundly, though, I took contextual dependence to be an ultimately unutterable metaphysical issue—as evident in my subsequent concern with (meta)physical deixis.[13] As such, I took context to be "pre-ontological"—more in line with continental discussions of "background" and "horizon" than with perspectives that felt it adequate to theorize context as an "extra parameter of meaning" to be incorporated into semantical equations—something I ended up caricaturing the "λcontext.content" view.[14] As the 1980s proceeded, my conviction grew that the latter approach, embodied in Barwise and Perry's situation semantics (cf. also McCarthy's 1996 "A Logical AI Approach to Context"[14.5]), was not just

11. This is not to suggest that self-directed reference is explanatorily or ontically prior to world-reference; as widely agreed, and explored in "Who's on Third? The Physical Bases of Consciousness," first, second, and third-person intentional capacities, singular and plural, undoubtedly arise coconstituted.

12. Barwise, Perry, Barbara Grosz (of sri International) and I were principal investigors on the grant which launched csli, a multi-institutional project spanning Stanford University, the Xerox Palo Alto Research Center (PARC), and SRI International. Long conversations about these issues figured in the formation of the Center's mandate, as well as of projects subsequently conducted underneath it.

13. See annotation A1, §2c ("Efficiency") on pp. ·12–·15, and o3.

14. Cf. annotation «where do I distinguish "meaning" and "content" in this way?»

14.5. «...Ref...»

excessively analytic but in fact missed the point. That is: though Barwise's and Perry's situation semantics approach did pay lip service to contextual dependence, I became increasingly convinced that, in virtue of remaining logicist or "formal,"[15] it had taken only a small first step towards making good on the original founding intuition—an intuition that I ultimately understood along something like the following lines: that, thrown into an ultimately ineffable world, we, as conscious beings, hew a roughly manageable ontological registration out of it—a registration profoundly dependent on a background or horizon of contextual enmeshing.

A6 ·6/0.7:9 Re the fundamental self-relativity of embodied organisms: cf. the discussion of deixis adumbrated in O3, as well as annotation A1.

A7 ·7/2 "You should think more of yourself, so that you can think less of yourself," I used to say to friends, in recommending psychotherapy. Whether others interested in reflective architectures and the technical intricacies of self-reference were saturated with echoes of psychoanalysis, Calvinism, and meditation practices I have no idea. But such considerations have never lain far below the surface of my work on the topics under consideration here.

A8 ·8/1/4:6 Disentangling notions of dependence and independence,[16] and opening up intermediate territories of partial (in)dependence, is a major theme in AOS, particularly Volume II, on Formal Symbol Manipulation; cf. also O3.

A9 ·8/1/11 In discussions in and around CSLI at the time (mid 1980s), John Perry was fond of using this phrase ("mental counties") for constituents of mental states. Whether it ever appeared in print I am not sure.

A10 ·9/1/4 By 'content' I mean something like referent or state of affairs that the expression is about (what in a more traditional philosophical context would be called "extensional content," and in the 3Lisp contexts of chs. 3–5 I called "declarative content"—signified by 'φ').

A11 ·9/1/4:6 Cf. the discussion in §6 of the Introduction.

A12 ·9/1/-7:-2 The claim that meaning is something that a representational structure "has on its own" is naive and misleading—though perhaps in

15. Cf. ch. 1, §... (pp. ...).

16. My concern was not with modal strength (nomological, ontological, metaphysical, logical varieties of indepdendence, etc.—distinctions that always struck me as more finely-chopped in philosophers' minds than in anything I could recognize as the world), but with what 'dependent' and 'independent'

this case mostly just infelicitously phrased. For one thing, no representational structure means on its own; whatever meaning it can be said to have, derivative or authentic, would derive only from its playing a role in a complex dynamic encompassing process. Moreover, the suggestion ignores myriad other issues, including for example the sorts of externalism about meaning that Putnam talks about in his example of 'beech' and 'elm'.[17]

The point of the paragraph was only to suggest the (rather simple) idea of meaning being something roughly like the "λ context . content" conception of meaning prevalent at CSLI at the time, as for example discussed in annotation A5, above.

A13 ·10/0/2:5 Cf. the discussion in «ref» of how, contrary to traditional practice in logic, I take denotation and proof-theoretic role to be semantic, not just the former—somewhat along the lines of the intuitions underlying "conceptual role semantics" in cognitive science and philosophy of mind. The approach derives from the strongly participatory thematic that underlies the critique of formality in AOS; cf. also discussions in "The Foundations of Computing, ch. 1.

A14 ·10/0 This paragraph is dense, and if useful at all, more so retrospectively than prospectively. I was clear that the myriad close relations between and among structures, processes, etc., were semantically implicated, and cognizant of the inadequacy of simple frameworks inherited from formal logic for dealing with the complexities of the computational case—but I by no means had adequate theoretical machinery with which to render that complexity intelligible.

A15 ·11/1/1 Specifically, the point made by Lewis Carroll in "What the Tortoise Said to Achilles,"[18] that modus ponens (deriving 'Q' from 'P' and 'P ⊃ Q') requires accepting an underlying (unwritten) premise or assumption of roughly the form '[P ⊃ [P ⊃ Q] ⊃ Q' which, if written down, requires accepting yet another underlying (unwritten) premise of the form '[P ⊃ [P ⊃ Q] ⊃ ⊃P ⊃ [P ⊃ Q] ⊃ Q] ⊃ Q', and so on and so forth, ad infinitum. No matter how much one attempts to make everything explicit, reasoning necessarily rests on accepting some additional inexorably implicit rules of procedure.

A16 ·11/1/9 The same point holds for the order of argument evaluation—whether a dialect evaluates its arguments in left-to-right or right-to-left order. Curiously, this fact is not even revealed by standard (non-

actually mean.

17. «Refs»

continuation-passing) meta-circular interpreter code, of the sort illustrated on p. 71 of the Lisp 1.5 Programmer's Manual (McCarthy et al. 1962), or p. 9 (fig. 2) of Steele & Sussman's "The Art of the Interpreter" (1978). Such interpreters are normally written in terms of a recursive EVLIS procedure, which (as usually defined) indicates that the implemented language processes arguments in the same order as the language in which the metacircular interpreter is itself written (a variant can easily be written so as to reverse this). The situation is different for continuation-passing interpreters, for reasons explained in [dissertation] §4.c.i ("Lambda Abstraction and Procedural Intension"), included here as ch. 3c.

A17 ·11/-1:12/0 These points foreshadow the general participatory theme that I was gradually developing when this was written. «…cf. FOC and Δ of boundaries …»

A18 ·13/1/5 As well as being a strain for organisms, dispensing with indexicality would be an even greater strain for theorists: we would have to describe the way people were structured each day, in a manner that took explicitly into account the metaphysical way in which each new day was different.

A19 ·13/1/-1 Note that this entire discussion—including Perry's—only gestures, and rather sketchily at that, at the ideas pointed at by the text, because it makes free use of such constructions as "same" and "different," both of which are defined with respect to types, which already build in many of the points being voiced here. To makes these points without some such presupposition is impossible; to say it as carefully as possible, while nevertheless acknowledging that ultimate limitation, would take many pages of complex metaphysics—and was not likely something I would have been capable of, at the time.

The issues tie into considerations of deixis, as discussed in §2c ("Efficiency") on pp. ·12–·15 (see also annotations A1 and A5; and 03).

A19.5 ·14/2/6 The term 'token-reflexive' is often used for essentially the same phenomenon. See. e.g., «…».

A20 ·14/-2 The arguments of location (38°N/120°W), orientation (187°N), vertical-orientation (GRAVITY-NORMAL), and time (3-JAN-86/12:40:04) held of the author at the moment the sentence was written.

A21 ·15/1/-4 By 'information,' in this sentence, is meant information content; cf. annotation A10, above.

A22 ·15/2/1:3 That representation involves non-effective (non-causal) relations to the distal is a theme that occupies much of my attention here and in subsequent work; cf. for example "The Representational Mandate," in "Rehabilitating Representation," ch. ... in Volume II; and 03.

A23 ·15//-2:-1 Issues of partial disconnection are a prominent theme of 03, q.v.

A24 ·16/1/-6:-3 If that representation of internal behavior is causally linked with how and what internal behavior actual comes about, the modification could then also take effect (come into being). This issue of causal connection, talked about on the next page (17/-1), was a major theme of 3Lisp and the reflective architectures explored in chs. 3–5.

A25 ·17/0/-5 Re something's mattering to a system or organism: cf. §1b.iii.α of ch. 3b (p. ...), and its accompanying annotation A30 (p. ...).

A26 ·17/-1/4:5 This notion of "appropriately connected detachment" was the issue for which the "causal connection" of 3Lisp (upwards and downwards reflection) was the design study. Cf. chs. 3–5.

A27 ·18/0/1:5 Tree rings might count as an example of totally connected representation; within the appropriate context (differential annual rainfall) they are fully connected to the lifetime of the live tree. And as this case illustrates, such representations are not necessarily useless, because of their (causal or computational) efficacy. That is: representations can represent, through the instantiation of effective properties, other properties or states of affairs that are not effective. Clearly, though, minimal representation of this sort does not provide an ability to entertain hypotheticals, sort into categories, be wrong, or serve as a mechanism via which to adjust the thereby represented facts.

A28 ·18/-1:19/0 As I would now say, using vocabulary introduced in 03, representations (or the systems or persons that employ them) register their subject matters—in terms of objects, properties, relations, states of affairs, etc., if they are what in the vocabulary of "The NonConceptual World"[19] I would call conceptual; in other ways if they are nonconceptual; but necessarily in some way or other. Cf. 03 chapters 7 and 7; also "Rehabilitating Representation."[20]

A29 ·20/-1/-2:-1 At least they must be local and effectively determinable within the

18. Caroll (1895).

19. Ch. ... in Volume II.

20. Ch. ... in Volume II

system as a whole; they need not be local and effectively determinable any more locally than that—i.e., at the level of the structure as an individual mereological ingredient.

A30 ·21/-1 To put this pedantically, one could say that immediacy is a relational higher-order property, since it has to do with the ability of (a tokening or occasioning of) another property to cause an effect; whereas syntactic, intrinsic, formal, etc., could be argued to be non-relational higher-order properties, if one felt that whether a property was or was not syntactic property depended solely on, as it were, 'local' or intrinsic facts about that property itself.

«...Do I believe this? Also: talk about its releation to (computational) effectiveness...»

A31 ·22/3/6 Here and elsewhere throughout my writings, it is my habit to generalise philosophically familiar notions of 'use' and 'mention' by extending 'mention' to apply to those objects referred to or named by uses of ground-level terms. Thus I would not only say (i) as would any philosopher, that in the sentence «The word 'Nile' contains four letters», the six-character expression «'Nile'» is used, whereas the four-letter expression «Nile» is thereby mentioned;[21] but also (ii) that in the sentence "The Nile is more than four thousand miles long," the four-letter expression «Nile» is used, and a very long river *is* mentioned. This extension not only accords with commonsense ("In recounting the incident, he mentioned his brother"), but has substantial philosophical pedigree—e.g., in P. F. Strawson's landmark "On Refering."[22] However the practice is not usually included within the use of the terms 'use' and 'mention' in technical discussions in logic and linguistic semantics.[22.5]

A32 ·23/1 This paragraph is a little wonky. In editing the paper for this Volume, I was tempted to improve or delete it in its entirety—but formulating an adequate theory of implicitness and explicitly (if such a thing could even be done) is far beyond the realm of editing; and simply to delete it seemed disingenuous. Particularly problematic is

21. When quoting expressions that themselves contain quotation marks, in these volumes I have used "French quotes"—i.e., the characters '«' and '»', for clarity. Cf. also footnote † on p. ·15 of ch. 4. Thus in the annotation French quotes are used as a mentioning device.

22. «Ref»

22.5. In computational discussions of semantics, the terms 'use' and 'mention' are not normally used at all.

its claim that "to be implicit is to play a role directly." While it is not difficult to discern some intuitions that might underlie such a statement, I am not currently willing to hazard a guess as to what exactly (if anything) I had in mind in writing it.

A33·27/1/-4:-3 By 'most of the self-referential mechanisms that have actually been proposed' I was referring to systems in AI that had been proposed as being self-referential—such as Weyhrauch's FOL, Moore's autoepistemic reasoners, etc. While it is clear from the text that, more generally, I would have counted any functioning email system to be a self-referential mechanism, those were not what I had in mind here.

A34·29/-1/5:6 It was exactly with the intent of developing such a "simple typology of relations of structured correspondence" that I wrote "The Correspondence Continuum," included here as ch. 12, which was presented during a conference that took place on May 21–23, 1986, almost exactly two months after this paper was written. Cf. §... of that chapter, where these notions of iconicity, objectification, absorption, and polarity are spelled out.

A35 ·30/0/1:2 The claim that "iconic relations are fully explicit" now strikes me as outright false. Whether it was a simple error at the time the paper was written, or an infelicitous statement of something insightful or even true, I do not currently know.

A36·30/0/-3:-1 Keys languishing in hotel mail slots is a dated example.There was a time, later, when the presence of voice mail on an answered telephone line indicated the absence of the person being called—another example I used to use—but it is a while, too, since that has been a reliable social regularity. Perhaps the most hauting example of a polar representation I know of is based on a rumour that was widely circulated in the 1980s, though I do not know whether it was ever validated: that, during the Cold War, if American submarines on patrol underneath the Northern polar ice cap did not receive an "everything is OK" signal from Washington for an entire week, they were on standing orders to come to the surface and launch a devastating barrage of nuclear weapons at the Soviet Union (the theory being that "no signal" was an adequate representation of a Soviet attack on American soil).

A37 ·31/B1&B2 «This needs to be thought through; I am not sure the characterization in the paper is correct—and if it is not, that deserves comment

here.»

A38 ·34//-2: 35//1 The statement that "B(α) should be represented just in case α is, and ¬B(α) just in case α is not" is ambiguous. On what might seem the more pragmatic reading, "just in case" should be interpreted as "only if," not as "if and only if"—i.e., that B(α), if represented at all, should be so only if α is. The "if and only if" reading, in contrast, seems at least on the surface problematic, because it generates an infinite number of introspective representations: if α is represented, then so too should be B(α), and thus so too should B(B(α)), and B(B(B(α))), etc., ad infinitum. Note, however, that, this threat of infinity notwithstanding, 3Lisp nevertheless embraced the "if and only if" reading—providing, albeit virtually, the full infinity of introspective representations. Programmers can use 3Lisp on the (simplifying) assumption that all such representations exist; as described in detail in ch. 5, the implementation always runs around in the background and puts them in place, before any use code can check on them, thereby allowing a finite implementation of the infinite ideality.

A39 ·35/1/1:4 To say that introspective integrity is not a property that a system must achieve is somewhat disingenuous. The point I was trying to make is that, though it should clearly be embraced as an ideal, it need not be adopted as an absolute requirement, since in some cases full introspective integrity transcends the bounds of finite computability, or is for some other reason impracticable That is: the hedge was included only in order to relieve the system (and its designer) from the need to achieve perfection. Nevertheless—as will be explored in considerably more depth in AOS—introspective (and reflective) integrity is a very substantial normative property: a characteristic in terms of which to judge systems, not merely an incidentally descriptive one.

A40 ·35/1/-2:-1 I have no idea what, if anything, this last sentence was trying to get at.

A41 ·35/-1/1 I.e., truth is not the only property that contributes to what I call the full significance of a sentence, statement, representation, or other intentional entity. Cf. the very first sentence of the abstract (·1/1/1), the discussions at 9/-1:10/0 and ff., the general significance function Σ introduced in regards to for 3Lisp in §1d.iii of ch. 3b (p. ·87),

23. Double for the following reason: «...explain...»

etc. In addition, needless to say, this is the meaning of 'significance' that is intended in the title of AOS.

A42 ·35//-1 «Ref»

A43 ·36/0/-10 «I.e., ψ=NORMALISE, as it were, in which the term on the left is in the theorist's external analytic language, and the term on the right is in 3Lisp's internal language.

A44 ·41/0/2 «Refs»

A45 ·41/n‡ In the original, the first sentence of this note read "At the time of its design I called 3Lisp 'reflective,' not 'introspective,' but I now think this was a mistake: reflection—see below—was what I wanted; introspection was what I had." However this is not quite correct. Cf. the discussion «where?» about how, though apparently self-referential in only an internal sense, it is to an extent a double consequence[23] of the knowledge representation hypothesis that the 3Lisp model of reflection has more to do with externally-focused reflection than might be evident from first appearances.

«...Check; note still reads that way? Figure out what is going on...»

A46 ·42//-3: 43//1 The requirement that a reflective system be able "to render explicit... the indexicality of its own embeddedness" is extremely important; cf. also the last sentence of the subsequent ¶ (43/1/-2:-1): that "reflection is necessary if one is to escape from the confines of self-relativity". «Possibly explain; also cite these points stuff in other discussions of the Δ between reflection and introspection.»

A47 ·44/0/2:3 One common way in which sentences and statements are distinguished, in philosophy, is to define a statement as a sentence that can be true or false (i.e., a sentence in indicative mood), as opposed to questions, commands, hopes, etc., which would be considered non-statement sentences. Sure enough, logic, focusing only on truth-evaluable sentences, deals only with "statements" in this sense. Since any sensible interpretation of commands, orders, questions, etc., requires adversion to a context of use, the passage in the text makes some sense on such a reading.

24. Even on this conception there are ambiguities as to whether, for example, "It is next to the bank" is one or two sentences, depending on which sense of 'bank' is intended; but none on whether, in an exchange in which two people utter "I really appreciate you" one or two sentences have been used—even if, as would normally be the case, in an utterance of the latter the (one) word 'you' were said more emphatically; the answer is one.

But this was not the distinction I had in mind. Rather, my point was more in line with the distinction drawn in Strawson's "On Refring" (1950), echoed in Barwise & Perry's situation semantics, according to which a sentence is understand as a grammatical type—an (abstract) sequence of words[24]—in distinction to (what Strawson consider) statements that can be made by using or uttering such sentences, in which the interpretations and referents have all been determined. Thus Strawson distinguishes uses of the sentence "The King of France is wise" made during the reigns of Louis XIV and Louis XV, claiming that in this example one sentence is used to make two statements (the first true, the second false). Evidently, contextual issues determine the difference between the two statements made. It is this sense of context to which I was referring (i.e., claiming that formal logic traditionally does not address).

A48·44/1/10:14 This conception of program is what I elsewhere call a specificational (as opposed to ingrediential) view; cf. «refs, once I figure out what I will use as the canonical description of this difference».

A49·46/-2/-2:-1 As noted in annotation A7, above, a vaguely Calvinist stance against self-indulgence and the contemporary privileging of "self"[25] permeates this and many other things I have written. In no way did I view my interest in reflection and self-reference as supportive of this trend; if anything, I was interested in discerning the genuine importance of self-reference (and self-knowledge) in order to rout the gratuitous excessive adversions to self that permeate the age.

The 20th century, I used to feel, would ultimately come to be regarded as the century of self-reference, and our fetishization of self historically seen in much the way we now view medieval obsessions with piety.

A50 ·47/-1/5 The Knights of the Lambda Calculus is a semi-mythical guild of 3Lisp programmers, afficionados, and fellow travelers, with headquarters famously located at the top of the 3Lisp reflective tower.

«Fix; cf. the following quote on Yash Tulsyan's site:[26]

> "A semi-mythical organization of wizardly LISP and Scheme hackers. The name refers to a mathematical formalism in-

25. Cf. the ubiquitous injunction to "enjoy yourself," enjoined in situations in which I have always felt it was much more interesting—to say nothing of more pleasurable—to enjoy others.

26. http://www.cosman246.com/jargon.html#Knights of the Lambda Calculus