

## **How I Learned to Stop Worrying and Love Reflexivity**

Jenann Ismael

Centre for Time, University of Sydney<sup>i</sup>

jtismael@u.arizona.edu

### **Introduction**

There is a lot of confusion surrounding reflexive thought, and strangely divided opinion about its significance. Confusions include that it is semantically mysterious, that it is incoherent, that it plays a role in the proof of Godel's theorem and allied results. Not everyone is subject to all of the confusions, but they're common enough to be worth dispelling. Some have held that the ability to think reflexively is distinctly human. Nozick, for one, regards it as the defining characteristic of a self. He writes;

“To be an I, a self, is to have the capacity for reflexive self-reference.”<sup>ii</sup>

For most of the major philosophers of the 17<sup>th</sup> and 18<sup>th</sup> centuries (not just Descartes, but Malebranche, Leibniz, Locke, Berkeley, and Hume), human cognition was understood as involving the mind's reflexive grasp of its own contents. Kant – an especially important and especially deep source of insight about reflexivity - regarded the capacity for reflexive thought and the capacity to conceive the world in objective terms as inseparable features of human cognition. They were, for him, two sides of a single coin.

But other important figures have thought that the very idea of a reflexive thought is incoherent. Ryle likened the idea of a reflexive thought to an arm that grasps itself. In another memorable image, he ridiculed the idea of a thought about itself as involving

“the hallowed paraoptical model, as a torch that illuminates itself by beams of its own light reflected from a mirror in its own insides.”

His suggestion was that reflexive thought doesn't make sense because mental representation, like grasping or illuminating, is a relation that one thing does to something distinct from it, not something a thing can do to itself. William James agreed with him asserting categorically that:

“No subjective state, whilst present, is its own object; its object is always something else.”<sup>iii</sup>

Ryle and James are not denying that one thought can be the object of another, they are denying that any thought can be its own object. The objection is based on a view about of the nature of mental representation, but others have been suspicious of reflexive thought because of unsavory association with paradox. If there are thoughts that are about themselves, then there are thoughts that assert their own falsity, and one might worry that allowing for the existence of reflexive thoughts opens the door for paradox.

Recent trends have focused on reflexive thought in a more positive way. Because reflexive thoughts share some of the semantic and epistemic peculiarities of conscious thought, it has become fashionable to suppose reflexivity holds the key to the mysteries of consciousness. Hofstaedter, D’Amasio, and many others, hold that conscious thought is by its nature reflexive.<sup>iv</sup> Kreigel and defenders of self-representational accounts of consciousness have proposed that to be a conscious state is to be one with a reflexive content.

But despite the currency of talk about reflexivity, it is hard to find an explicit account in the contemporary literature of what a reflexive thought is. One sometimes hears that a reflexive thought is one that *is about, refers to, or represents* itself.<sup>v</sup> This purely extensional definition, however, fails to capture the difference between (A) and (B) below.

(A) A contains four words.

(B) I contain four words.

They are both about themselves, but the latter form a special subclass with quite special semantic and epistemic properties, and it is these that we mean by ‘reflexive’.

Among contemporary authors, John Perry has written the most about reflexivity.<sup>vi</sup> He gives a very elegant account of the cognitive and semantic peculiarity of indexicals in terms of reflexive content. Reflexive content, for Perry is a layer of content independent of the ordinary referential content.<sup>vii</sup> It is *truth-conditional*, but it is distinguished from the referential content by the fact that reflexive content specifies the truth conditions of an utterance (or thought) in terms that make reflexive reference to the utterance or thought itself. So, for example, consider the following utterance made at noon on 01/16/09

(C) “Today is John Perry’s 63<sup>rd</sup> birthday.”

The referential content of (C) is the singular proposition that on 01/16/09 is John Perry’s 63<sup>rd</sup> birthday. The reflexive content of (C), by contrast, is that the day on which (C), itself occurs is John Perry’s 63<sup>rd</sup> birthday.<sup>viii</sup> In general, reflexive content on Perry’s account relates the subject matter of a sentence, utterance, or inscription to the sentence, utterance, or inscription itself. Perry is persuasive in arguing that there is a quite special kind of content that attaches to reflexive utterances and thoughts, but in conveying the truth conditions for reflexive utterances, he makes use of the notion of ‘the utterance itself’, which is itself a reflexive notion,

and so for the purposes of understanding the difference between reflexive and non-reflexive thoughts it is not quite enough.<sup>ix</sup> What follows is an attempt to give an explicit semantic characterization of specifically that difference.

### **Representation of self**

Representation will be conceived for our purposes as a relation,  $xRy$ , between the elements in some representational medium  $X=\{x_1, \dots, x_n\}$  and objects in the domain of representation  $Y=\{y_1, \dots, y_n\}$ .  $X$  might consist of terms in a language and  $Y$  physical objects, or  $X$  might consist of a stock of symbols on a map and  $Y$  geographical objects like cities, highways, rivers, and so on.  $R$  is given by a model-theoretic mapping of  $X$  into  $Y$ .  $X$  is used to convey information about  $Y$  by arranging  $\{x_1, \dots, x_n\}$  into configurations that reflect the structure of  $Y$  under the intended mapping.

In the general case,  $R$  is an external relation, so that:<sup>x</sup>

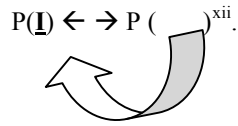
- (i) There are no constraints on internal relations between elements in  $X$  and elements in  $Y$ ,
- (ii) There are no constraints on spatial and temporal relations elements in  $X$  and elements in  $Y$
- (iii) For any pairing of elements in  $X$  and  $Y$  such that  $xRy$ , each of  $x$  and  $y$  can exist without the other.

Words, for example, don't have to resemble or be located near what they represent. They just have to be arrangeable into sentences that convey information about physical states of affairs. Dots and lines on maps don't have to resemble cities and highways, they just have to be arrangeable on paper to reflect spatial arrangements in the world. Properties (i)-(iii) are crucial to the practical point of representation. We have a use for easily produced, portable proxies that allow us to talk about things that don't exist, or to exchange information when they are not present. They are what provide the advantages of saying over showing.

Most of the time, when one is talking about representation, one assumes that  $X$  and  $Y$  are disjoint, one assumes, that is to say, that the items doing the representing don't themselves fall into the scope of representation. But this assumption is not necessary, and indeed it fails for some of the most important representational media. It fails for thought, it fails for natural languages, and indeed for any medium with unrestricted scope. And in the cases in which it does fail, as a direct consequence of its failure, it is possible to construct representations that are about themselves. So, for example, in the on the top right corner of my desk, I have inscribed ; « The inscription in the top right corner of ---'s desk contains thirteen words». And I might say (or think) as the clock hands pass midnight on the eve of my 40th birthday « the first thought of my 40th year will be a reflexive one ». <sup>xi</sup>

These inscriptions, utterances, or thoughts are *of*, or *about*, themselves in a purely extensional sense.

But once we have a medium with tokenings of expressions that represent themselves extensionally, we can extend it by introducing the expression “**I**” with the rule that: an inscription/utterance/thought of the form “P(**I**)” is true just in case P is true of *it*, i.e., the inscription/utterance/thought in which it is contained.



Call this **I**-extended language  $L^*$ .  $L$  contains sentences A, and B from

**D**                    *The sentence inscribed in the top right corner of \_\_\_'s desk contains fourteen words.*

**E**                    *E contains four words*

But it does not contain F

**F I** contain four words.<sup>xiii</sup>

For every sentence in  $L^*$  that can be expressed using **I**, there is a counterpart in  $L$  with the same truth conditions, obtained by substituting a coreferential name or definite description for the relevant occurrence of **I**.<sup>xiv</sup> In that sense, the expressive power of  $L^*$  doesn't outrun that of  $L$ .

The difference between reflexive expressions and non-reflexive singular terms is the generally licensed substitution of **I** for the sentence in which it is contained, a substitution that requires no empirical knowledge and is licensed in any linguistic context or context of use. Consider, by way of contrast, the non-reflexive singular term which is the subject term in “The sentence inscribed in the top right corner of \_\_\_'s desk contains fourteen words”. Let's call the singular term ‘The sentence inscribed in the top right corner of \_\_\_'s desk’ G and the sentence as a whole, D. G, as I said, in fact refers to D, but the fact that the substitution of G for D is truth preserving depends on empirical facts that aren't generally available in the context of use. Unless you happen to be in my office looking at my desk in good light, with knowledge that it is my desk that you're looking at, there are at least epistemically possible worlds in which that substitution is invalid. And G can also occur in linguistic contexts in which it doesn't refer to the containing sentence, so the substitution of the sentence on display in the context of use for G can't be built into the semantic rules that govern the use of

G.<sup>xv</sup> For an example of an (epistemically) possible world in which D is false, imagine the world in which I've inscribed in the relevant place on my desk "Constructive Empiricism Rocks". For an example of a linguistic context in which G can occur but can't be substituted for the sentence on display, consider the sentence "The sentence inscribed in the top right corner of ----'s desk contains three swear words." And because of this, the inference from "*The inscription in the top right corner of ----'s desk contains fourteen words*" to the manifestly (or 'inspectibly') true "The inscription in the top right corner of ----'s desk contains fourteen words" is not licensed by the semantic rules governing the expression. Inspecting the inscription will not tell us it is true. In the overwhelming majority of contexts in which D occurs, there will be epistemically possible worlds in which it is false. And in the overwhelming majority of linguistic contexts in which G occurs, the substitution of the sentence in which it is contained is not truth-preserving.<sup>xvi</sup> These two features in conjunction are what distinguish "I" from coextensive definite descriptions (including rigidified definite descriptions), on the one hand, and names, on the other.<sup>xvii</sup> Reflexive thoughts generate a new epistemic category of truths; not necessary truths, or *a priori* truths, but contingent truths that are known to be true by inspection; ('truth by inspection'). I always, on any occasion of use, in any linguistic context, refers to the thought, inscription, or utterance in which it occurs.<sup>xviii</sup> And that means the substitution of something that is *guaranteed to be exhibited or displayed in the context of use* for I— namely, the thought/utterance/or inscription itself—is a part of the fixed semantic profile of the I. It's part of the rules governing the use of I that one is allowed to make that substitution and judge for himself, as it were, whether what is *asserted* matches what is being shown.<sup>xix</sup>

It might be thought to share this latter feature with quotation. Quotation is a mechanism for generating names for expressions by enclosing an expression in a special syntactic markers. The word cat becomes "cat", dog becomes "dog", C above becomes "the sentence in the top right corner of ----'s desk contains thirteen words", and so on. Because a quoted expression is syntactically a part of its name, it shares with reflexive representation the feature that it displays an expression as a way of representing it. The difference between quotation and reflexive representation is that a quoted expression is represented when it is displayed, but not used. It functions as a name for itself, but it does not refer to or assert anything when it is enclosed in quotes. An I-containing sentence, by contrast, is simultaneously used *and* displayed. This means that although one can construct a sentence that says something about a sub-sentential component using quotes, one cannot construct a sentence about itself using quotes. One can try, but the sentence will appear inside quotation marks in the place of a singular term, leaving the predicative part of the sentence outside quotes). The same will go for other media and larger units of representation. A picture can't contain a precise, accurate, non-reflexive reproduction of itself that reproduces it in full detail as a proper part. And an individual consciousness can't contain a precise, accurate, non-reflexive copy of its all of its contents and still have room to represent anything else.

### **Epistemic and semantic implications**

There are epistemic and semantic guarantees that follow from the validity of that substitution that are unique to reflexive representation:

- Uses of **I** are guaranteed a referent,
- A competent user of **I** is guaranteed to be able to identify the referent *de re*.
- Since an **I**-containing inscription (or utterance)<sup>xx</sup> presents itself for inspection at the same time that, and with the very act in which, it represents itself *as* thus and so, it permits a direct comparison between what is *said* and what is *shown*. An inscription of “**I** am written in red”, or an utterance of “**I** am said in a booming voice” is, if true, *self-evidently* so.<sup>xxi</sup>

None of these things are true when representation proceeds by way of an intermediary, because an intermediary can always exist without the existence of its referent and need not share the properties of what it represents. They arise because representation reduces to presentation in the reflexive case, and it inherits all of the epistemic properties of presentation.

Let me caution, however, against the implication that because it comes with these guarantees, reflexive representation is an especially *good* type of representation. On the contrary, reflexive representation is a degenerate form of representation.<sup>xxii</sup> These guarantees can’t be secured without undermining the practical point of representation. In the ordinary case the link between a thought and its truthmaker is an external one, a conventional association embodied in a model-theoretic mapping that imposes no constraints on relations between the vehicle and subject matter of representation.<sup>xxiii</sup> The externality of that link is crucial to how representations are ordinarily used. It is what allows us to agree on a simple, transportable proxy that can stand in for an object when we want to talk about it, something that we can produce at will and that will let us exchange information about the object when it is not present. A representation that is inseparable from what it represents has none of the advantages of saying over showing. If one had to display a thing in order to talk about it, arms would be strong and conversation limited.

### **The Reference-Grounding Problem**<sup>xxiv</sup>

One way of bringing out what is special about reflexive representations is by seeing them as an answer to a problem that emerged from a discussion that Putnam started back in 1975 when he presented the original version the Model Theoretic Argument. The argument spawned a massive literature. What the argument *is* and what it really shows remain matters of ongoing controversy. But a revisionary version of the argument due to David Lewis highlights a role that only a reflexive construction can play grounding reference.<sup>xxv</sup> Formal models usually assume the disjointness of X and Y. Languages are treated as abstract objects interpreted by a model-theoretic mapping from terms into elements in its domain. The argument proceeds as follows. We start with a level of non-semantic fact W (for ‘world’) and a first-order language  $L_0$

that is used to represent how things are in W. To represent the relations between  $L_0$  and W, we introduce a richer language  $L_1$  that has separate terms referring to elements in each. The richer language provides a side-on perspective that allows us to compare different ways of mapping  $L_0$  into W. That can't be done in  $L_0$  itself, because using  $L_0$  to pick out elements of W, presupposes a mapping of  $L_0$  and W. If we try to represent that mapping in  $L_0$ , we get the uninformative definition "A" refers to A, "B" refers to B, and so on.<sup>xxvi</sup>  $L_1$  likewise represents things in its domain (which includes both  $L_0$  and W) but not its own relations to that domain. Those relations, in their turn, are given in a richer language that has names for expressions of  $L_1$  as well as those in its domain. In this way, a hierarchy of increasingly rich languages can be constructed, each interpreting languages below it in the hierarchy.<sup>xxvii</sup>

$L_n$   
 ...  
 $L_2$   
 $L_1$  language  
 $L_0$   
 W non-linguistic fact

This hierarchical picture is deeply entrenched and embodied in the model-theoretic apparatus that is used to model representation of all kinds; linguistic, mental, pictorial, diagrammatic. The Model-theoretic Argument begins with the assumption that if there is a language at the top of this hierarchy – a language that doesn't get its intended interpretation from a definition in yet another higher level language – it must be thought. After all, there's nothing outside of the head that interprets our thoughts for us, and we ourselves don't have a side-on view of the relations between thought and the world. Putnam argued us that unless we can fix the intended interpretation of the mental symbols that encode our theory of the world before we try to use the symbols to state the theory, the only way to come up with a theory that will be constrained enough to be false is by making it logically inconsistent or incompatible with the world's cardinality. The reason is that we'll be simultaneously asserting the theory and using it to pick out its intended interpretation – effectively saying 'here is my theory, and the intended interpretation of the language is one that makes it come out true'. And since there will always be at least one an interpretation that makes in come out true (barring inconsistency or cardinality mistakes), it won't be false.<sup>xxviii</sup>

The claim that we have to simultaneously use the theory to implicitly define its own interpretation is based on the claim that the mind has no other way of identifying an intended interpretation, because, as Putnam says.

“the mind never compares an image or word with an object, but only with other images, words, beliefs, judgments, etc... On any theory, when the child learns the use of the word ‘table’, what happens is that the word is linked in certain complex ways (‘associations’) to certain mental phenomena’<sup>xxix</sup>

If one could identify the intended interpretation by pairing words up directly with what they represent, one could say that the intended interpretation is the one that assigns ‘dog’ to [putting in here not an image or word or other form of representation of a dog, but a dog itself]’. Thoughts like this – like thoughts in an interpreted metalanguage - would bridge the gap between thoughts and things. But without a direct comparison, and without the side-on view provided by an interpreted metalanguage, no attempt to ground reference for one’s own ideas from an internal perspective by framing thoughts of the form “‘dog’ refers to dogs” will have any success. Let us say a representational system is *closed* just in case it has no representationally unmediated access to the domain of representation. And let us say reference is *grounded* just in case there is some way of establishing relations between its terms and elements in the domain of representation. This line of argument suggests that there is no way for a closed representational system to ground the reference of its own terms. Putnam is not denying that we possess quite elaborate theories – theories of mind and theories of reference - that represent the relations between thoughts and things, but if Putnam is right, the interpretation of those theories is determined by the very facts that they are meant to convey. The attempt to represent how our ideas relate to reality, is like an attempt to tell ourselves what our words mean using our very own words. It is at best a way of organizing the internal space of representations; it might constrain relations among ideas, but it won’t tell us how the whole body of internal representation is related to anything external.

This is a modern way of putting the old and familiar complaint that the mind is trapped in the circle of representation, able to link representations with one another but unable to break out of the circle and connect any representation directly with a thing itself. It is a worry about whether any constrained, informative internal representation of the relation between thought and the world can be formed based on the claim that the interpretation of any such representation is going to depend on (be determined by) the very relation it is trying to depict. This is obvious for definitions of the form “‘dog’ means dog, where we are using the very same expression on the left and right. It is less obvious, but still true, on Putnam’s view, of apparently substantive theories of mind and reference, so long as those theories are part of our theory of the world, and the intended interpretation of our total theory of the world is whichever one makes it true. The argument for this is what came to be called the ‘just more theory’ response, which was deployed most famously against those who proposed a causal theory of representation. One proposes a substantive theory of representation;  $x$  represents  $y$  just in case  $x$  bears an certain relation  $C(x,y)$  to  $y$ . The problem with this is that this attempt to fix the meaning of ‘ $C(x,y)$ ’ suffers the same fate as the rest of the theory. Without independent constraints on the interpretation of the vocabulary used to state the theory, one ends up simultaneously asserting the theory and using it to implicitly define its own interpretation, and ‘ $C(x,y)$ ’ will end up referring to that relation, whatever it is, that makes the theory true, and no progress has been made. The same objection will work against any

theory of representation that requires interpreted vocabulary to state, so long as that interpretation consists of a mapping between internal and external elements.

An analogy can be used to convey the difficulty. Suppose Ann wants to paint an abstract or symbolic portrait of a family, one that represents members of the family but doesn't resemble them in the standard way. She seats them in front of her and produces a framed portrait containing a striped circle, a lopsided oval, a tiny star, and a triangle. The relation between the shaped blotches of paint and the people seated in front of the canvas isn't itself depicted in the portrait. That's an external relationship between the canvas and the seated group. It would be easy for Ann to convey her representational intent to an onlooker with whom she shares a side-on visual perspective on the canvas and its subjects. But what about someone who could only see what was inside the frame, is there some way for Ann to convey her representational intent to such a viewer? On Putnam's view, that is the situation the mind is in; trapped within the frame, without a side-on view of its own relation to the world.<sup>xxx</sup>

Ann could paint another picture, a meta-painting, perhaps making use of the original by embedding it in a larger frame in which it appears as a painting, painting a new image of the family in front of it and maybe drawing arrows from shapes on the embedded painting to images in front. Whether she employs a new representational scheme in the meta-painting, or the same representational scheme as the old, the same questions that arose original elements would arise about her pictorial depiction in the larger frame. She can go on, embedding the new picture in a larger frame and drawing arrows to depict its relationship to its intended subject matter, and so on... reproducing in pictorial form, the ascending hierarchy of languages from above. But at no stage will she manage to paint something inside the canvas that will ground the representational intent of the whole hierarchy by connecting one of the symbols inside the frame to a real bit of breathing flesh. And this suggests that any representation of the relations between mind and world is going to be *ungrounded from an internal perspective* unless one can establish some direct, representationally unmediated connections between mental representations and what they represent.

Discussion of the Model-Theoretic Argument in the literature has tended to focus on the question of what *determines* interpretation, and the most influential responses have rejected Putnam's insistence that "we determine interpretation or nothing does", holding that the facts that determine the what our ideas and internal representations represent are (partly) external to the mind. So, for example, causal theorists hold that it is the web of causal relations between ideas and the environment that ground the interpretation of ideas, not the fact that we have beliefs in our belief boxes of the form "x represents y just in case C(x,y)". David Lewis holds that there are divisions built into nature that constrain which mappings are candidates for an intended interpretation. But this leaves entirely open the separate, and equally interesting question of whether interpretation is internally grounded. One can hold that the facts that *determine* reference are external to the mind, and still raise the question of whether we have any constrained mental grasp of the facts that determine interpretation. Is it true, as the argument suggests, that the mind (or any closed representational system) is

confined to the quite precisely uninformative mental equivalent of thoughts of the form “‘dog’ refers to dogs”, “‘snow is white’ is true *iff* if snow is white”, and so on in any attempt to grasp what ideas represent?

### **Reflexivity and Reference-Grounding**

Putnam was certainly right that *ascending* semantically is an ineffective way of grounding a representation of the relationship between mind and world, because the interpretation of the representation would depend on the very facts it was supposed to represent.<sup>xxx1</sup> Is there another way for a closed representational system to provide an internal grounding for the terms that it employs to represent? Yes, a closed representational system employing medium that includes elements like the reflexive “I” above can ground the interpretation of at least some ideas by linking them in reflexive thoughts with internal representations themselves. Reflexive thoughts, utterances, or inscriptions allow the mind to associate elements in an interconnected hierarchy of mental representation with objects or events that are connected in the domain of representation, items that have a place among the non-linguistic furniture of the world. The semantic rule that defines the use of **I** when it was introduced above allows us to move from content of an inscription, utterance, or thought, conceived as an internal mental act, to the inscription, utterance, or act itself. Reflexive thoughts thus serve as semantic anchors constraining relations between the whole web of ideas to the domain of representation. Enough semantic anchors in the right setting can in principle determine a unique interpretation. “**I**” provides a counterpart of semantic ascent, allowing passage between semantic levels in the downward direction. Where devices like quotation allow a system to pass from a representation into a representation of a representation, reflexive terms allow a system to pass from a representation to what it represents. A thought of the form “**I** occur at t”, for example, links the temporal term “t” with an event, rather than just a representation of an event. A thought of the form “**I** belong to the person who is spilling sugar all over the floor” associates the description “a thought belonging to the person spilling sugar all over the floor” with a thought rather than a representation of a thought. All acts of self-location and self-identification have a reflexive content and the reflexive content anchors or grounds the interpretation of the non-reflexive content. Relating this back to Perry’s account of reflexive content, reflexive content on Perry’s account relates the subject matter of a sentence (its referential content) to the utterance itself. We see here now what that means, and we see the semantic importance of the links reflexive contents establish between referential content and non-linguistic reality.

But what about the ‘just more theory’ objection? Couldn’t the argument be run against the proposal here? How does one know what “**I**” means? No. The challenge presented by the MTA is a worry about whether the mind could place constraints on the relation between internal representations and external things, i.e., about any inner act of intending could reach outside the mind and hook the word ‘dog’ up with the furry four-legged creatures we know by that name. The rule of use for “**I**” doesn’t require any leaping outside

the head. There is no ‘meaning’ to know here that requires one to fix a relation to an element in the external domain. “**I**” adds a new rule of transition that places constraints on interpretation that entirely concern relations between internal elements. That’s the genius of it.

### **Why the suspicion of reflexivity?**

So why the suspicion of reflexive thought? One might speculate that there are several reasons.

- They don’t conform well to a Fregean Model of thought, and more generally, philosophy is too entrenched in the representationalism that leads us to expect that every component of thought should have the job of *standing for* something. A more pragmatic approach to semantics – one that focuses on what the parts of a thought *do* rather than what they *represent* - will be better equipped to recognize how devices like “**I**” function semantically.

- The popularity of perceptual models for cognition, suggests reflexive thought is impossible. So we get, for instance, strained metaphors like Wittgensteinian analogy of the seeing eye that falls always outside the field of vision, the Rylean metaphor mentioned earlier of the reaching arm forever out of reach, the camera futilely trying to capture itself in its field of vision, the sentence trying to quote itself, and always falling partly outside the scope of the quotation marks, and so on.

- Visual representations like the Escher drawing above suggest that there is some trick or ambiguity that is illegitimately exploited.

- The most serious, and the one that hasn’t been addressed yet, is worries about paradox.

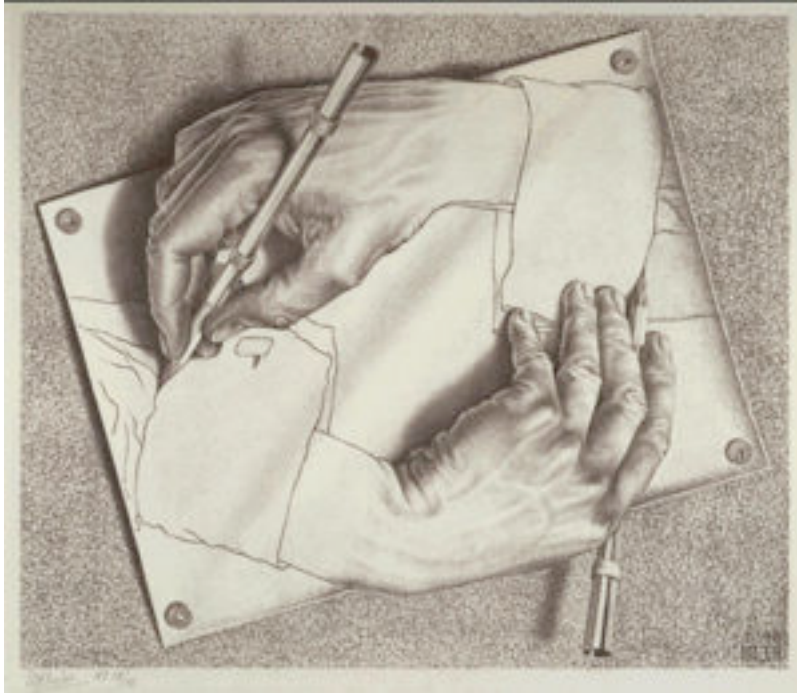
With the exception of Descartes’ “Cogito ergo sum”, probably the most famous reflexive sentence, at least among philosophers is Godel’s “I [i.e., this sentence] is false”. The question of whether adding **I** to a medium generates inconsistency needs to be addressed. The introduction of **I** is a conservative extension. It only makes explicit a potential for paradox that arises when you add semantic predicates – ‘is true’ or ‘refers to’ – are added to a language and allowed to apply to its own expressions. One doesn’t need reflexivity to generate paradoxes, only a truth predicate that applies to utterances, thoughts, or inscriptions in which it occurs. Although they’re frequently paraphrased using reflexive expressions, the sentences Godel uses in his Incompleteness proof do not contain reflexive expressions.<sup>xxxii</sup> They use numbers assigned as names to the sentences in which they occur by a Godel numbering. They don’t have the form “**I** am false”, but “n is false” where ‘n’ is a singular term with a fixed semantic value. Kripke showed how to generate paradoxes with attenuated semantic loops without any appearance of reflexivity.<sup>xxxiii</sup> Even the original Cretan – the reported source of the liar paradox – didn’t use reflexive expressions. He just claimed that all Cretans are liars. So adding reflexive expressions to a language is not necessary for generating paradox. Nor is it sufficient. One can avoid paradox even where there are reflexive thoughts by avoiding talk of truth or by restricting the extension of semantic predicates so that they don’t apply to the sentences in which they occur.

On the question of how I introduced by the rule relates to the token-reflexive that Reichenbach introduced in 1947, and other indexicals there are many positions one could take. <sup>xxxiv</sup> My inclination is to think the most basic is the generic ‘this very’ that combines with concepts (this very thought, this very utterance, this very day, this very place...) to yield reference. Indexicals like ‘today’, ‘here’, ‘now’ are easily analyzed in this way. I’m also inclined (with less assurance) to think that the I of the individual thought is more basic than the token-reflexive, in the sense that it is the minimal unit of self-reference. But in all cases, we only have the idea ‘this very X’ when we have the idea of an X. So, for example, the child that has the idea of place, but not yet a clear and distinct conception of thought, can have the idea of ‘this very place’, but not ‘this very thought’.

There’s a connection with Descartes that is worth pointing out here. The inference from “I think” to “I am” in the cogito argument is legitimate so long as the I of the *cogito* is the reflexive I of the individual thought. The only thing whose existence is inferred on this reconstruction is the thought itself. And this would accord with Descartes’ insistence that in asking what the self is he is asking “what is this I whose existence is made known to me in the very act of trying to deny that it exists?”. The substantive step in the argument comes, on this reading, not in inference to the existence of the self, but in bridging the gap between the I of the individual thought and the “I” of the extended consciousness. For this latter bridges multiple thoughts. <sup>xxxv</sup>

### **Level-shifting and tangled hierarchies**

What is special about reflexive representations can be explicitly stated now. They cross semantic levels. They provide linguistically licensed opportunities to pass between semantic levels in the downward direction. <sup>xxxvi</sup> A thought of the form “P(I)” takes us from the representational content of the thought, to the thinking event itself, which is connected in space and time and has a place in the causal fabric of the universe. I-containing thoughts have a sort of semantic ambiguity that is exploited to let us pass from the space of words to the domain of things, and back again. There is a very famous illegitimate shift between representational levels in the Escher picture Drawing Hands that has a disorienting effect on the viewer.



Here there is the level of the picture and the level of reality that it represents, and an unlicensed shift between representational levels that creates an impossible situation. Reflexive expressions provide perfectly legitimate, linguistically licensed versions of this sort of shift. Another example of a legitimate shift that formally has the same structure as a reflexive utterance, and which does better in some respects as a model for a reflexive consciousness, is that of a map with a red dot. In representational terms, the red dot represents the map in which it is contained. In terms of the rules that govern inference and substitution for map symbols, there is a linguistically shift down a representational level, from the red dot in the representational space of the map to the map itself. And just as in the case of the reflexive sentence, this association between an item in a representational space and an object rooted in physical landscape does what the version of the Model-Theoretic argument above suggested was impossible, creating a semantic bridge between the representational content of the map and the physical landscape in which the map is situated. To generalize this example to something that might begin to provide a model for the reflexive consciousness, the featureless red dot needs to be replaced with an internal image that represents not only the location of the map, but its shape and size. A reflexive mapping into the image will, then, ground representations of spatial distance and orientation. And if the internal image also represents the distribution of colors over the map's surface, a reflexive mapping will ground representations of those colors. A colored map equipped with an internal self-image drawn to scale equipped with a reflexive mapping will tell the viewer what represents distances and colors in the representational scheme of the map by presenting her with instances of spatial relations and colors and mapping them onto properties and relations in the internal self-image. If the map is one foot long from top corner to top corner and a viewer wants to know what represents one foot in the representational scheme of the map, she checks how long the internal self-image of the map is, corner to corner. If the map has a red

patch on the left side and a green patch on the right and a viewer wants to know what represents red and green in the representational scheme of the map, she checks how those patches are represented in the internal self-image.<sup>xxxvii</sup> And once she has grounded representations of red and green, she has grounded representations of any properties that can be identified by their relations to red and green.

This strategy for the interpretation of descriptive vocabulary is effectively the same as it is for spatial vocabulary. Once one has grounded reference of some location terms she can use these to ground reference to others by setting up a system of coordinates that implicitly relates other points to those whose reference is grounded. And likewise, once one has grounded the reference of some property terms she can use these to ground reference to others by using them as points of reference with respect to which the others can be located. The strategy, applied to a medium with more qualitative richness will provide one with a correspondingly richer basis of grounded property representations, and these in their turn can serve as the basis for identification of a wider circle of properties. I have argued elsewhere that experience is best thought of in these terms, as having a reflexive content that relates the properties it exemplifies to its subject matter. Others have held similar views. The status of experience is a complex and hotly contested subject, and an explicit account of the nature of reflexive content is needed to assess such views.<sup>xxxviii</sup> The account of reflexivity given here can be put to service in that capacity.<sup>xxxix</sup>

### **Conclusion for reflexivity?**

There are good reasons that reflexivity has been the source of contention in philosophical circles. Reflexive expressions don't conform to our ordinary models of representation and without a good understanding of their structure, their seemingly magical properties rightly makes them a source of suspicion. The mystery is dispelled by providing an explicit semantic characterization of reflexive thought that explains and underwrites some of their very special properties.

---

<sup>i</sup> I am grateful to audiences at Princeton University and the Work in Progress Seminar at the Centre for Time.

<sup>ii</sup> *Philosophical Explanations*, Harvard University Press, 1981.

<sup>iii</sup> William James, 'On Some Omissions of Introspective Psychology', *Mind* 33, 1884, p. 2.

<sup>iv</sup> For a collection that draws together a broad range of discussions of self-representational approaches to consciousness, Uriah Kriegel and Kenneth Williford, *Self-Representational Approaches to Consciousness*. MIT Press/Bradford Books, 2006 and Uriah Kriegel, *Subjective Consciousness: A Self-Representational Theory*, Oxford University Press, 2009. See also Gilbert Harman, "Self-reflexive thoughts." *Philosophical Issues*, 16 (2006): 334-45.

<sup>v</sup> *Wikipedia*, for example, defines 'self-representation' as: a reflexive representation 'a sentence or formula that refers to itself via some intermediary expression or encoding.' See also Kriegel (2007), *ibid*.

<sup>vi</sup> John Perry, *Reference and Reflexivity*, CSLI Publications, 2001,

<sup>vii</sup> The referential content (which Perry also refers to as the "official content"), on his account, is given by the ordinary Fregean truth conditions.

<sup>viii</sup> That this is different from the singular proposition that 01/16/09 is clear from the fact that one could believe the singular proposition and not realize that (C) itself occurs on that day, and likewise, one could know that (C) itself occurs on Perry's 63<sup>rd</sup> birthday without knowing what date that is.

<sup>ix</sup> To see that this use of 'the utterance itself' presupposes the distinction between reflexive and non-reflexive characterizations of the utterance, it is enough to notice that the reflexive content wouldn't be captured if "the utterance itself" were replaced by "the utterance made by JI at noon on Jan. 16, 2009". In some places, Perry

---

says is that to have a reflexive thought to have a thought whose truth conditions make reference to that very thought, ‘under the guise of identity’. But it’s hard to understand what is meant here.

<sup>x</sup>Some have argued that R is not a two –place relation between a term, or element in a representational medium, but a three place relation with a suppressed contextual parameter. Perhaps, but this won’t affect points here.

<sup>xi</sup> For simplicity, I’m treating thought here as a form of subvocalized utterance.

<sup>xii</sup> I’m making some choices here, assuming that the basic unit of self-representation is the sentence and the basic form of the self-representational sentence is ‘P(I)’, i.e., ‘I have property P’. We might instead have taken the basic unit of self-representation as the singular term. In that case, we simply hold that reflexive sentences of the form “I have property P” should be analyzed as ‘the sentence that contains this token has property P’. Reasons for thinking the I of the individual thought is more basic than the token reflexive I include reasons for thinking that thoughts as wholes are more basic units of reference than their parts. It may be that the first personal “I” of the self-attribution (“I believe that p”, “I think that q”) is more basic, because it may be that the “I” only makes sense in the context of the reflexive consciousness. I believe that this is the argument in Kant. See Beatrice Longuenesse “Kant’s “I think” versus Descartes’ “I am a thing that thinks”” (ms.) attributing arguments effect to Kant.

<sup>xiii</sup> Note that the I here is different from the first personal I of ordinary English, which refers to the person who utters it, and also from the token-reflexive I of Reichenbach, which doesn’t refer to the sentence in which it is contained, but only to the part of the sentence that it constitutes. See note \_\_\_ below. See my \_\_\_ for an account of the first-personal I.

<sup>xiv</sup> That will keep the truth conditions fixed, but change the truth value in cases like ‘I contain five words’.

<sup>xv</sup> I assume a fixed interpretation, unless otherwise specified.

<sup>xvi</sup> Another example; consider

**D** Sentence 23 is \_\_\_’s (single most) favorite sentence.

**D\*** “Sentence 23 is \_\_\_’s (single most) favorite sentence” is \_\_\_’s favorite sentence.

**E** *The inscription in the top right corner of \_\_\_’s desk* is \_\_\_’s (single most) favorite sentence.

**E\*** “*The inscription in the top right corner of \_\_\_’s desk* is \_\_\_’s (single most) favorite sentence” is \_\_\_’s favorite sentence.

If D and E are true, D\* and E\* are false (indeed, D and E are true *iff* D\* and E\* are false).

<sup>xvii</sup> I’m not going to be fussy about distinguishing uses of “**I**” from mentions. I’ll suppress quotation marks unless there is unclarity in the context.

<sup>xviii</sup> The individuation conditions for the referent depend on the individuation conditions for occurrences in a sense that depends on the medium of representation. Inscriptions are treated as objects, and their individuation conditions depend on spatial location. Utterances are treated as events and their individuation conditions depend on when they are uttered and by whom. Thoughts, likewise, depend on the identity of the thinker and time.

<sup>xix</sup> An **I**-containing sentence *presents itself as what is represented*.

<sup>xx</sup> I’ll suppress the bracketed phrase for ease.

<sup>xxi</sup> Or, perhaps, self-evidently *true to users that can, respectively, see and hear*. Note that this doesn’t mean that, except in special cases, they have to be true.

<sup>xxii</sup> In mathematics, a special case of a relation in which arguments that are usually distinct coincide is said to be degenerate. Degenerate cases are limiting cases in which a relation reduces to a different, usually simpler class. So, for example, since the roots of an nth degree polynomial are usually distinct, the two identical roots of the second-order polynomial make it a degenerate case.

<sup>xxiii</sup> I use ‘subject matter’, following Perry, to mean ‘referential content’, or non-reflexive truth conditions.

<sup>xxiv</sup> What follows is an abridgment and clarification of some of what I said about the reflexive response to the model-theoretic argument in SS.

<sup>xxv</sup> David Lewis, "New Work for a Theory of Universals", *Australasian Journal of Philosophy* 61 (1983), pp 343-77), "Putnam's Paradox", *Australasian Journal of Philosophy* 62 (1984), pp 221-36.

<sup>xxvi</sup> This is the Tarski definition of truth for Model-Theoretic languages. There are allied definitions for reference. Tarski, A. 1944, "The semantic conception of truth", *Philosophy and Phenomenological Research* 4, 13-47.

---

<sup>xxvii</sup> All languages above are assumed to be first-order. This hierarchy can be reproduced within a single language if the language is not first order by equipping it resources for semantic ascent. Restricting the presentation to first order languages highlights distinctions between semantic levels.

<sup>xxviii</sup> Perhaps this is better put in negative terms, since there may be many different interpretations that make it come out true. We should say, without an independent way of identifying the intended interpretation, it will not come out false. The burden is really on the opposition to show how interpretation can be sufficiently constrained to allow it to come out false.

<sup>xxix</sup> Hilary Putnam, *Realism and Reason. Philosophical Papers*, vol. 3. Cambridge: Cambridge University Press, 1983. viii., viii.

<sup>xxx</sup> Except that Putnam wants to eliminate the analogue of a painter outside the frame whose representational intent determines the interpretation of thought. That is the point of his repeated insistence that “we interpret our languages or nothing does.” Ibid., p. xii.

<sup>xxxi</sup> Precisely because “‘Dog’ refers to dogs” is true no matter how ‘dog’ is interpreted, believing “‘dog’ refers to dogs’ doesn’t tell us anything about how ‘dog’ is to be interpreted.

<sup>xxxii</sup> The proof is given in a first-order language without an analogue of **I**.

<sup>xxxiii</sup> So, for example, Alice thinks that Bob is thinking something true Bob thinks that Alice is thinking something false, so that what Alice is thinking is true *iff* what it is false

<sup>xxxiv</sup> Hans Reichenbach, *Elements of Symbolic Logic* (New York: Macmillan, 1947), 284ff. For criticism of the token-reflexive theory, see Quentin Smith, “The Impossibility of Token-Reflexive Analyses”, *Dialogue: Canadian Philosophical Review*, Vol. 25, No. 4, Winter 1986, pp. 757-760. “Sentences About Time”, *The Philosophical Quarterly* 37 (1987), 37-53.

<sup>xxxv</sup> Anscombe also suggests this reading, and argues that the argument founders on its inability to bridge the gap between the I of the individual thought, and the extended “I” of the first person.

<sup>xxxvi</sup> A thought of the form “I occur at t” the temporal term t with an event, rather than just a representation of an event. A thought of the form “I am the sort of image, smell, sound, or touch causally associated with such and such” associates the description “the sort of image, smell, sound or touch causally associated with such and such” with an image, smell, sound or touch, rather than a representation of one. It has been argued – I would so argue – that getting the logic of experience right requires attributing experiences this sort of reflexive content.<sup>xxxvi</sup> A thought of the form “I belong to (am an event in the psychological history of) so and so” associates the description “an event in the psychological history of) so and so” with an event rather than just a representation of one. These thoughts are also sometimes called ‘self-locating thoughts’, and Bas has a lot to say in his new book about the way in which self-location (ostension and measurement) links elements in scientific models with the systems they represent. He deals only with the interpretation of external artifacts and in his account, model gets related to the world in the situated pragmatics of application. In his story, the user is an ineliminable interface between model and world. We would like a generalization of this story that didn’t commit us to the existence of a homunculus playing the role of user, relating thoughts to the world. The reflexive account gives us that. It give us a reflexive account that gets rid of the middle-man.

<sup>xxxvii</sup> Note, that the map needn’t be the identity. We might just as well let green represent red and red represent green. Or we might use numbers to represent them. Any one-one function will do.

<sup>xxxviii</sup> How do we apply this schema to mental representation? Concepts of phenomenal properties are grounded reflexively, and other concepts are grounded by causal links to phenomenal profiles. See ---. This requires that the causal relation is itself exemplified internally so that the concept of cause is grounded reflexively. That is something I accept. This strategy will not work unless there is some relation that spans the space between mind and world and that is itself internally exemplified. It works for the interpretation of maps because maps are spatially extended; the spatial relation plays that role.

<sup>xxxix</sup> See ---.