Knowledge as Personal: The Representation of Self in the Representation of Knowledge

Joel Parthemore

Paics Research Group
Department of Informatics
University of Sussex
Falmer, Brighton UK

Abstract

I begin with a broad discussion of the nature of knowledge, leaving aside characterizations of knowledge as justified true belief in favor of divisions into conceptual/non-conceptual and *knowing that/knowing how*. Knowledge is always knowledge of: knowledge requires an object, but not an object on its own, for the object requires a subject. Just as one probably cannot begin to understand conceptual knowledge without an appreciation of non-conceptual knowledge, so, too, one cannot understand knowledge, conceptual or non-conceptual, in the absence of an understanding of intelligence. Some philosophers have argued that any understanding of intelligence needs to be broad-based and not anthropocentric; but I will argue that what I call the "anthropocentric stance" is at least useful and at most possibly necessary. This anthrocentric stance is fundamentally a part of our relations to other human intelligences, and it may be fundamental to our relations to *any* intelligent entities. In understanding others, we begin with an understanding of ourselves. In understanding *other* intelligences, we begin with an understanding of *human* intelligence. In understanding knowledge, "self" and "other" are continually recurring concepts.

Along with the anthropocentric stance, I present a version of the "meta-level argument" showing how representations can be organized into a hierarchy of representations, representations of representations, and so on. Extending a line of thought from the anthropocentric stance, I suggest that our representation of "other" begins with a modified representation of "self". In turn the representation of "self" may begin with a representation of boundary: the "self" from the "not self". The meta-level argument provides a way to talk about three distinct but related notions of self, which I call "I_1", "I_2", and I_3. I will argue that these are not simply distinct notions of self but are related in a hierarchy of first-, second- and third-order representations. The "self" can masquerade as the homunculus in the mind without being one, as Daniel Dennett has pointed out. This is not a paper about consciousness, but it does attempt to show how a certain approach to consciousness might shape any subsequent approach to knowledge representation.

1 The Nature of Knowledge

Before there ever were professional philosophers to ask them, there were the questions: What is it to know? What is it to be known?

Knowledge, says the definition, is understanding. Knowledge is always knowledge *of*. The framework on which it is hung, be it representational or non-representational, not only has no usefulness but has no meaning on its own. Be it *knowing that* or *knowing how*, knowledge requires an object: knowledge is always *about* something. But framework and object on their own still are not enough, for the object requires a subject; "knowledge", in any sense in which we may wish to use it, is always knowledge possessed or used – or by some arguments at least *potentially* possessed or used – whether by someone or something. What would a library, the repository of *knowing that*, be without any readers or patrons? What would any skill, the domain of *knowing how*, be without the skilled? The idea of any understanding of knowledge completely abstracted away from an agent possessing or using that knowledge is probably incoherent. Knowledge does not form a closed system unto itself.

Knowledge is probably not a natural kind. What it is depends on how one views it and how one uses it. It is not of the world but a useful abstraction from the world: a model of the world that, relative to some agent, makes the world more approachable. At the same time, knowledge of the world is all we ever have of the world, for we never have the world itself, unmediated.

Knowing how shows, for some, that knowledge must involve more than conceptual knowledge... certainly does, if concepts are taken to be those units of knowledge that form the components of explicitly propositional thought and meet the Generality Constraint: they can be re-used, re-combined, recycled in a systematic way. If I can think about new bicycle frames and white laptops, I can think about white bicycle frames and new laptops, without those concepts losing or changing their core nature along the way. I can even, if I have in mind touring bicycle frames and Macintosh laptops, entertain such curiosities as Macintosh bicycle frames and touring laptops and attempt to assign some meaning to each.

But even if we take an agnostic view on the "explicitly propositional" requirement or deny it outright, the Generality Constraint would seem to require room – quite a bit of it – for non-conceptual knowledge: for surely perception involves a great deal of transitory knowledge that is not only not consciously accessible, it is not re-usable, for it is never kept for re-use or at least never gets conceptually structured: it is and remains "raw data". If the Generality Constraint is used as the dividing line, then conceptual and non-conceptual knowledge form figure and ground: understanding and representing the first requires an attempt to understand and represent the latter. The distinction between *knowing that* and *knowing how* may significantly overlap with that between conceptual and non-conceptual, but there may well be instances of *knowing how* that do meet the Generality Constraint, that are conceptually structured but not consciously accessible, as there may be instances of *knowing that* that on the Generality Constraint test, fail to be conceptual. All of this, of course, allows us to remain agnostic on the question of whether conceptual and non-conceptual knowledge share a common form, whether they can be built up using a common schema.

2 The Anthropocentric Stance

If knowledge of object requires knowledge by subject, then something needs said about the relationship between knowledge, subject and object, which is intelligence: that is, the ability of a subject (or "agent") to apply knowledge. Philosophers and cognitive scientists can debate all they like whether or not, when they are considering whether or how to represent knowledge, they are modeling that knowledge (*knowing that* or *knowing how*, conceptual or non-conceptual) in the same way that a human agent does[1]. But I believe that there is no way, when we talk about "knowledge representation", that this knowledge can fail to be knowledge *in a human sense,* which is to say, from a human point of view; for when we seek to understand intelligent agents, our models for that understanding are human agents (as when we seek to model other human agents, I will argue, our models are ourselves).

Such a human-centered approach to understanding intelligence is, I believe, precisely the sort of thing Alan Turing had in mind in his 1950 paper, *Computing Machinery and Intelligence:* roughly, if we consistently treat an artifact (like a computer) as intelligent – what Daniel Dennett calls taking the Intentional Stance toward it – then it makes little practical sense to argue whether or not it actually *is* intelligent. Turing describes an "imitation game" with three players: two human and one an artifact. It is the goal of the one human player, communicating only through the written word, to guess which of the other two is the human and which the computer. It is part of the game's design that the nature of the players' embodiment is not, *in and of itself*, allowed to prejudice the outcome against the computer.

> The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include. We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against

[1] E.g.: "We must be careful to distinguish the question of whether such and such a program constitutes a good model of *human* intelligence from the question of whether the program... displays some kind of *real, but perhaps nonhuman* form of intelligence." [Clark 2001, p. 20]

> an aeroplane. The conditions of our game make these disabilities irrelevant. The 'witnesses' can brag, if they consider it advisable, as much as they please about their charms, strength or heroism, but the interrogator cannot demand practical demonstrations. [Turing 1950]

Blay Whitby has famously argued that the "Turing test" has led AI research down a "blind alley" by being interpreted as encouraging, if not actually encouraging (i.e., by intention of the author), "an operational test for intelligence involving some sort of comparison with human beings". [Whitby 1997] Whitby is far from the only researcher who has claimed that a human-centered approach to understanding intelligence – what I will call here the anthropocentric stance – is a serious mistake; but he has expressed his views in a particularly clear way and at some length. Human intelligence is a poor starting point because it is too poorly understood and excludes too much that we would want to include as intelligence.

> Unfortunately AI scientists, rather like early astronomers, tended to look at the subject from their own perspective. That is to say they saw human intelligence as the starting point and wanted to develop the scientific study of intelligence from this starting point. [Whitby 2003]

This to me begs the question: what other perspective should they – or could they even – have used? The view of a subject is, by definition, subjective – and subjective from not just any viewpoint, but specifically that of the subject in question. Of course their perspectives were limited, in ways that in retrospect we can find quite amusing; but what grounds do we have for claiming that our own perspectives are not, on the whole if not in this one particular area, just as limited? Those who postulated an earth-centered universe were *wrong*; their theories eventually fell to Occam's razor[2]. But this doesn't mean that their theories weren't a useful, even a necessary, stepping stone to the theories we have today, which may themselves be only a stepping stone to the theories we hold in future. There is a longstanding tradition in science which holds that science is *not* about establishing any objective truths about the universe but rather about proving existing theories wrong (i.e., inconsistent or incoherent) or simply less preferable than simpler theories that seem to explain the observable facts as well or better.

Of course Whitby is right that "science has to be interested in the whole space of intelligence". (Whitby, 2003) It might seem arrogant indeed to assume that intelligence must be human or human-like (humanoid, if you will) intelligence... to define intelligence so that only human intelligence (as we currently understand it) fits the bill.[3] But what precisely is wrong with *starting* from a human-centered understanding of intelligence Whitby never makes clear. Is it not at least as reasonable to argue that, as our understanding of intelligence expands to include other species or artifacts, that our understanding of what it means to be human – to have human intelligence – will not likewise expand? As the Marc Almond song goes, "tell me if you can / what makes a man a man".

3 Levels and Meta-Levels

A lot of arguments can be made clearer, and a lot of apparently unresolvable problems (particularly of the Cartesian dualist kind) can be avoided, by some form of the meta-level argument: simply put, that what seems an issue or a problem at *one level* of analysis may be examined away from a meta-level of analysis[4]. This is the distinction to be made between a self-referential paradox and a simple contradiction. The paradox's apparent contradiction (between its sense and its [self-]reference) can be resolved by use of a meta-level argument: "I am lying" can be analyzed in such a way that the sense of the proposition is at one level, the [self-]reference at the meta-level. This allows the self-referential paradox to be interestingly

[2] Theoretically, there's no reason why one couldn't, *in principle*, revive Tycho Brahe's system of epicycles on epicycles to explain not only planetary motion but also all of the space exploration that has taken place to date. The problem is that any such system would be horrendously complicated, to the point of being farcical, compared to much simpler explanations that explain the observable facts without constantly needing to add more epicycles.

[3] ...Though as Ron Chrisley has pointed out (personal communication), writers like Hubert Dreyfus[Dreyfus 1992] could be read as taking this view. Likewise Wittgenstein can be read as saying that we would be unable to recognize non-human intelligence as intelligence: "if a lion could talk, we could not understand him." [Wittgenstein 2002]

[4] Blay Whitby first raised this point for me.

meaningful in a way that a simple contradiction like "Tuesday is Wednesday" or "day is night" (where the sense of the one item excludes the sense of the other). But what precisely do we mean when we talk about a "meta-level argument"?

The idea of different levels of description or levels of analysis should be familiar enough: by any one level, we mean a certain context or perspective; by any other level, either an abstraction away from that level or a level that *that* level is itself an abstraction away from. In one direction there is a qualitative loss of information, in the other direction a qualitative increase. At the one end of the spectrum is the physical world; the concrete; the essence of raw, uninterpreted data. At the other end is the metaphysical (or meta-metaphysical) world; the abstract; the essence of high-level, interpreted, processed and categorized data. So "different levels" can generally be read as "different levels toward or from abstraction". A meta-level is always relative to some "level" and is an abstraction away from that level. In this way we can talk about objects in the physical world (base level) or (e.g., mental) representations of those objects (meta-level) or representations of those representations (meta-meta-level), and so on. We can talk about thoughts (base level) or thoughts about thoughts (meta-level) or thoughts about thoughts about thoughts (meta-meta-level).

So for an example from the realm of neuropsychiatry and human agents, here is Joseph LeDoux, talking about different levels in the mind and their apparent independence and interdependence:

> The conscious and unconscious aspects of thought are sometimes described in terms of parallel functions. Consciousness seems to do things serially, more or less one at a time, whereas the unconscious mind, being composed of many different systems, seems to work more or less in parallel. Some cognitive scientists have suggested that consciousness involves a limited-capacity serial processor that sits at the top of the cognitive hierarchy above a variety of special-purpose processors that are organized in parallel.... (LeDoux, p. 280)

For comparison, here is an example from the realm of AI and artificial agents, Terry Winograd talking about different levels in a computer program and focusing on their independence:

> ...For a typical complex computer program, there is no intelligible correspondence between operations at distant levels. (Winograd and Flores, p. 90)

I've discussed why what I call the "anthropocentric stance" is not only *not* an obstacle to (conceptual) knowledge representation but may actually be quite useful, even essential. I've presented a version of the meta-level argument as a means to ordering the representations one creates. I want to use the anthropocentric stance to argue for a certain sort of relationship between our mental representations of "self" and "other". I then want to use the meta-level argument to argue that our different notions of self (self as physical organism, self as mental creature, self as name or set of descriptions) can usefully be arranged into a hierarchy of representations and representations of representations.

4 The Other as Self

It seems a definitional part of being a human intelligence that our understanding of intelligence and of self begins at the center, as it were, and grows outward: think of the child first aware of itself, then its mother, then its father, then of siblings and other family members, and over time more and more examples of like-me-but-not-me's. It seems counterintuitive – though of course by no means ruled out! – to think that intelligence could work the other way around: a conception of others preceding a conception of self, self as the destination rather than the starting point. (Still: what would it mean to have a concept of other people and *not* of ourselves?)

I want to suggest that our understanding of another person begins with seeing that person as being somehow like us.[5] All other things equal and until we are given reason to believe otherwise, we expect the other person

[5] Indeed to treat someone as fundamentally *not* like ourselves is arguably a necessary step to treating her as less than a person, even sub-human. Consider the nature of racism or the treatment of enemies in war. Because there is such a close

to be motivated by the same things that motivate us; we expect the other person's thoughts to follow the same sorts of paths that ours do. One of the explanations offered for autism is the failure to make this link. Consider: when I reflect on the Christmas gifts I buy for people, I realize I tend to buy the sorts of things that I would like myself. When I "feel" for someone who is homeless or unemployed, it is because I can picture what it would be like for me to be in the same position.

When we fail to understand the other's motivations – through the perspective of our own – when their thoughts fail to follow the expected channels, we experience a break down: the script has been departed from. What is this strange someone who I thought was like me and isn't? It is when others are very apparently different from us – often in particular visually, as in the case of someone with a gross physical deformity – that we frequently experience the greatest discomfort.

The imposition of ourselves onto the not-ourselves needn't stop there, of course, and arguably doesn't: we are constantly imposing ourselves onto the world. We are constantly re-creating the world in our own image.

But this is all high-level, socially deeply embedded discussion; so let us try to simplify things by returning to the earlier discussion. If knowledge requires intelligence, then intelligence requires mind (as mind would seem to require body). Mind may or may not require representations, but let us consider for sake of argument that it does.

What do we mean by "representations", a term that seemingly can be used in so many different ways? I suggest we use the provisional definition "complex and potentially reusable constructions of themselves reusable symbols." Likewise symbols we might define as "atomic entities that function to 'stand in place of' some complex structure (i.e., representation), and that can be manipulated according to some formal rule or rules, such that any semantics interpretable by an agent to the system are preserved." (What others would call a "complex symbol" would, in this scheme, be called a representation.) Some researchers would, of course, like to deny the symbols and keep the representations (using some more general definition of the term)[6]: that is to say, analyzing representations into any compositional rule-governed structure is, for them, either a fruitless or an impossible task. Others would like to deny the representations as well: Tim van Gelder sees them as dispensable[Van Gelder 1998] and Randy Brooks is well-known for seeing them simply as a hindrance[Chrisley 2003].

I want to suggest that mind, be it self-conscious, subconscious or sub-personal, can be viewed at an abstract, functional level as a network of interconnected and inter-defining representations. (This is as opposed to the familiar image of independent, context-free representations.) I want to suggest further that our familiar concepts of "self" and "other" are among those mental representations, and that our representation of other as like-me-but-not-me might well begin with a copy of our mental representation of ourselves, subsequently perturbed – reshaped – by information and experience. My representation of you, in other words, begins with my representation of me. Picture a figure of malleable clay, or the science fiction stories of clones-as-perfect-copies (as opposed to "real" clones, which despite superficial resemblances are *never* perfect copies): from the moment of their cloning they begin to diverge, to the point that they may eventually become (superficially, if not inherently) unrecognizable to the original.

Of course from the moment we meet someone we may use various cues (skin color, apparent ethnicity, accent, physical deformities) to represent them mentally in terms of *other* people we have met in ways that, on the surface at least, may be quite unlike ourselves. But these differences belie, I think, more fundamental similarities. In representing others, we rarely if ever wander far from the original pattern.

---

connection between our sense of self and our sense of our humanity, to treat someone as fundamentally *not* like ourselves probably is to make them something *other* than human.

[6] This is the position that Fodor and Pylyshin, accurately I think, attribute to most connectionists[Fodor and Pylyshin 1993].

5 The Self as Other (I_1)

> *It must be me because I'm here.* That is what Emily said cautiously as she contemplated the face in the mirror before her. It had to be her; she had placed herself in front of the mirror, of her own free will, so it had to be her; who else could it be? (Damasio, p. 162)

In **THE FEELING OF WHAT HAPPENS**, Antonio Damasio describes a woman who is unable to recognize herself in a mirror. In most other ways she functions like any other "normal" fully conscious human being; but her connection to unique visual images, including her own image, has been lost. At one level, her sense of self is functioning perfectly. At another level, it is not. She both knows and does not know herself.

If we mentally represent other people as somehow-copies of our representations of ourselves, then it is reflexively true that at some level our representations of ourselves must be like our representations of others. What is more subtle is that if we use our mental representations of ourselves to create our mental representations of others, then it is also likely to be the case that we use our representations of others to create – or re-create, or modify – our representations of ourselves. In this way the "odd" notion I mentioned earlier of starting one's understanding of self on the "outside" and working in toward the "center" has a grain of truth. It is an illusion, if a useful and possibly necessary one, that we can ever perceive ourselves directly, for that-which-we-are. There is a truth in the old idea that other people are the mirrors by which we view ourselves. Take the metaphor of a web: there is a sense in which each of us is the center point of her own web; there is another sense in which each of us is just another point in the web, defined by all of the points around us.

This is the first of three notions of self-as-mental-representation that I want to consider in this paper. I will call it I_1. It is a first-order representation: the first abstraction away from the original. This first self is the self-as-other, the "third-person" representation of self that is like but unlike all the many other representation of selves: distinguished from the rest because it applies to *this* self. It is, in some peculiar way, *self-similar* to the "self" – the organism, if you like – in which it resides, fitting it like a hand into a glove (only, we should not mistake the hand for the glove). It is self-referential in a way that none of our our other mental representations are, for it references the *entire* system in which those representations reside.

The $I_1$is a representation of the physical organism as a whole. But because we have no direct access to that physical organism as a whole – our access to it as to any external entity is mediated through our senses – what we standardly treat as the actual organism is, in fact, a representation of that organism.

There is an essential continuity to the represented: human beings are not "shape shifters", that science fiction creation that can take whatever form of embodiment it chooses. There is likewise an important continuity to the representation: it changes, but only within narrow parameters. Note the use of the indexical. "I" is an indexical, as is "this", as is "you", as is something like "today". "I" references whichever agent is using it at the time. Just as when I use the word "you", or talk about "today" or "this" moment, the reference is determined by the immediate context. And yet there is a sense I stubbornly wish to hold onto in which, when I am talking about *this* "I", which is me, I have a continuity in mind that the other indexicals lack, something that is more than the continuity of one day following another or, to echo Damasio, one self following another[7].This indexical, in short, is special.

6 The Self as "Myself" (I_2)

The "core self", for Damasio, is created and destroyed in each moment. Continuity enters at the level of the "autobiographical self", which brings all the moment-by-moment core selves together, telling a story that unites all the core selves all in a common narrative. As such it is a representation of representations: that

[7] Discussing the "core self": "Just as death and life cycles reconstruct the organism and its parts according to a plan, the brain reconstructs the sense of self moment by moment." [Damasio 2000, p. 144]

is, a second-order representation, one step further removed from the physical organism. This is the self to which are attached various stable facts: a name, a place of birth, a nationality, an occupation, and so on.

> The organization of consciousness I propose resolves the apparent paradox identified by William James – that the self in our stream of consciousness changes continuously as it moves forward in time, even as we retain a sense that the self remains the same while our existence continues. [Damasio 2000, p. 217]

Damasio's core self is a low-level representation, close to the physical level. The self-as-other I describe above is a higher-level (that is, more abstracted) representation, and as such affords a greater appearance of continuity. What they have in common is that they are both first-order representations: a representation of the organism as it exists, or is perceived to exist, in the world.

What precisely is the first-order representation of self and what the second-order representation depends a bit on one's interests: Damasio is a neuropsychiatrist, seeking to find the biological structures that underlie consciousness; he is interested in locating the first-order representation as close to the physical level as possible. My self-as-other would, in Damasio's approach, lie somewhere between the core self and the autobiographical self. What is important is not so much where one divides things but in the notion of levels and meta-levels, of representations and representations of representations: one sense of self abstracted away from another sense of self abstracted away from the physical organism. This is the idea I want to take away from Damasio.

The second thing to note is that, although Damasio's autobiographical self is clearly a second-order representation, a representation of the core self representations, still most of his discussion of the core self and autobiographical self occurs along a continuum: most of the time, he does not refer to core self and autobiographical self as discrete levels. My interest in this paper precisely is in discrete levels of self. The situation is analogous to discussions of physical dimensions, which can be considered either as discrete levels (two dimensions define a plane, three dimensions define a volume, and so on) or as existing along a continuum: so-called fractal or Hausdorff dimensions[8] .

So Damasio's core self is both related to, and distinguishable from, the self-as-other I have described. The autobiographical self is, on the other hand, very close to the next sense of self I want to talk about: what I will call the "self-as-'myself"' or I_2, itself a second-order representation. This is the self-reflective self: the self that has a sense of itself. This, Damasio suggests, is what humans have and other animals don't.

The I_1 is a physical organism. The I_2 is a mental creature. The I_2 is, if we are careful not to confuse the metaphor with the reality, the homunculus sitting in his Cartesian theatre of the mind, controlling the shell of an organism in which he sits and observing all that it observes (as another homunculus must be observing him, and another, and another!) Ideally, we would like to keep the image and ditch the ever-regressing homunculi. The mental creature is, I want to suggest, a fiction (being "merely" a representation of a representation of the physical organism), but an extremely useful fiction: one that makes the organism much more flexible in its responses to its environment.

Normally we conflate the I_1 and the I_2 – or we finesse between them. It may in fact be necessary that we do so. It would be strange indeed, in most cases, to think of one person having two (or more) quite distinct notions of self. But sometimes the two representations of self, first-order and second-order, get separated in some important way, as with Emily. She has a core self; she has an autobiographical self. But some of the links between the two are broken. In particular, she (being the autobiographical self of extended consciousness) cannot visually recognize herself (a mental image that is part, or would normally be part, of her core self). But she can recognize herself in other ways: listening to her own voice, or touching herself.

[8] Discussion of levels and meta-levels, I would like to suggest, is simply a generalization from the concept of physical dimensions. In the one direction, toward the abstract (fewer dimensions, or lower fractal dimension), there is a qualitative loss of information. In the other direction, toward the concrete (higher dimensions, or higher fractal dimension), there is a qualitative increase of information.

So the disconnect is selective. Note that, from Damasio's clinical experience, the autobiographical self – or what I am calling the I_2 – is usually quite aware of this partial disconnect:

> Not only is she conscious of what she knows perfectly well, but she is also conscious of what she does not know.... Emily, as well as the many other patients like her that I have studied over the years, is perfectly conscious of the things she does *not* know and she examines those things, in reference to her knowing self, in the same way she examines the things she does know. [Damasio 2000, p. 163]

What Emily's case suggests to me is what seems the natural progression of the train of thought I've been developing: namely, that even in ordinary circumstances we only have direct access to the I_2, not the I_1. The I_1 mostly if not entirely resides in the subconscious: that part of the conscious mind that is *not* the self-conscious. Our access to the I_1 is always mediated through the I_2. When breakdowns in this mediation occur, we find cases like Emily's. Like it or not, our access even to ourselves is always through a mind's "I" view.

If the I_1 is already an abstraction away from the underlying reality – by which I mean that there has been a qualitative and quantitative loss of information from the original to the representation – then the I_2 is an abstraction of an abstraction. Each time we step further away from the reality – creating a useful abstraction that makes that reality easier to understand or relate to[9] – something important is lost.

Some time ago I took part in a generative art project called Chinese Whispers, organized by a Brighton (UK) artist. Although it is certainly far from a precise analogy, I would like to draw a comparison. In this project, volunteers (who might or might not be artists) were asked to sketch a copy of an original line drawing. Then a second set of volunteers copied those copies, a third set copied the second-generation copies, and so on. Although there was, on the face of it, no requirement for either qualitative or quantitative loss of information from one generation of copies to the next (as there is in my description of I_1 and I_2), nonetheless after only a few generations (never more, I think, than four or five), the sense of the original drawings was entirely lost on any observer[10].

7 The Self Impoverished (I_3)

To my two notions of "I" I have given so far, I_1 and I_2, I would like to add and briefly mention a third: I_3. This would be I_2's understanding (or representation) of itself. Who does the "I" who thinks "I" think that "I" is? Since I_2 is already a representation of a representation (or abstraction of an abstraction), that makes I_3 a third-order representation.

While we normally think of the mental entity (the I_2) *doing* things – making decisions, coming to conclusions, effecting changes to the physical organism (the I_1) – I_3 more or less just sits there, an important but not very entertaining place holder. Call it the name by which I know myself. A collection of some descriptions (including one's own name), it is subject to revision and some degree of manipulation by I_2, but that is all.

Beyond this point, there seems nothing to be gained in going further. Why? Why does it not make sense for I_3 to have a representation of itself (I_4) – an even more severely impoverished sense of self, to be sure, but a self all the same – for I_4 to have a representation of itself (I_5), and so on? Why could there not be an infinite series of selves beyond selves reminiscent of Douglas Hofstadter's talk of "enlightenments yon enlightenment"? [Hofstadter 1979, pp. 231-245] Why is there no regress?

[9] Compare what the neuropsychologist V.S. Ramachandran has to say: "...Your concept of a single 'I' or 'self' inhabiting your brain may be simply an illusion – albeit one that allows you to organize your life more efficiently, gives you a sense of purpose and helps you interact with others." (Ramachandran, p. 84)

[10] More information about the artist and the project can be found at http://www.rachelcohen.co.uk.

I_1 exists and moves in a physical world, as perceived through our senses. I_2 exists and moves in a mental world that is abstracted away from the physical world. But where can I_3 be said to exist and move save in some abstraction away from the mental world to which we can give no name? We can imagine it transcribed into something like a folder in a filing cabinet, and nothing more; we cannot attribute it any volition. Where could we even begin to locate I_4? Not only can we attribute it no volition, we have no substance to give it.

The difference between the endless regress of homunculi and my series of "pseudo-homunculi" I_1, I_2 and I_3 that seem to "bottom out" at the I_3, is that the former is ineliminably dualist, while the latter offers an escape. The former locates all the homunculi at the same level (of complexity or abstraction); the latter separates them out into level, meta-level and meta-meta-level. There is no infinite regress ("of course", one might exclaim) because each homunculus is qualitatively and quantitatively simplified from the last. As Dennett points out as early as BRAINSTORMS [Dennett 1981], we only get an infinite regress when we assume at each level a homunculus equal in abilities, in complexity, to its predecessor. Compare an actual person, a pencil drawing of that person and a textual description of that drawing (and so, indirectly, of the person). With each step toward abstraction, a huge amount of information is lost.

8 Conclusions

Concepts exist independently of neither subjects nor objects, so that a representational system modeling that knowledge must take account of both the subject that possesses those concepts and the objects to which those concepts refer, showing the relationships between all these levels. One way to take account of the subject is to model it internally to the representational system. At the same time this provides a natural way to discuss mental representations of self.

Fears about the pitfalls of taking an anthropocentric stance toward understanding intelligence seem, on the face of it, unwarranted, or at least answerable. The perspective offered by the anthropocentric stance may be painfully limited, but it may also be the best that we have, and the limitations may not be all that they seem. Indeed, they may prove to be a benefit, by giving us a needed starting point. The representation of subject (or "self") in a natural or an artificial intelligent system may well serve as the model for representations of other intelligent agents.

There is an important sense in which we appear constantly to be re-creating the world in our own image; our understanding of the minds of others may well begin with our understanding of ourselves. In similar fashion, our initial representations of others may well be copied from our representations of ourselves. Our representations of others can then be used to offer feedback to and help modify our representations of ourselves.

As the anthropocentric stance offers guidance to how we might approach our mental representations of "self" and "other" and their relationship to each other, so, too, a form of the meta-level argument allows us to relate in orderly fashion our different competing notions of self, showing how they can be arranged in a hierarchy from most concrete (the physical organism itself) to most abstract and why we might want to consider them in that light.

The end goal in doing knowledge representation is not understanding knowledge, conceptual or otherwise, as some abstract, disembodied entity. It is to better understand ourselves and what knowledge means for us. The lesson for AI that I would like to go away with has been expressed succinctly by Ron Chrisley:

> Perhaps here (finally) we have a reason why AI must be made in our own image.... Because that is the only way that we will be able to grasp and refine the concepts necessary for AI development. If this is right, giving our AI systems a robust form of embodiment may have as much to do with developing our own mental abilities as it does with developing theirs. (Chrisley, p. 148)

References

[Brooks 1991] Brooks, Rodney A. (1991). *Intelligence Without Representation*, **Artificial Intelligence** (47), pp. 139-159.

[Chrisley 2003] Chrisley, Ron (2003). *Embodied Artificial Intelligence,* **Artificial Intelligence** (149), pp. 131-150.

[Clark 2001] Clark, Andy (2001). **MINDWARE: An Introduction to the Philosophy of Cognitive Science**. Oxford University Press, Oxford.

[Damasio 2000] Damasio, Antonio (2000). **THE FEELING OF WHAT HAPPENS: Body, Emotion and the Making of Consciousness**. Vintage (Random House), London.

[Dennett 1981] Dennett, Daniel (1981). **BRAINSTORMS: Philosophical Essays on Mind and Psychology**. Prentice Hall, London.

[Dreyfus 1992] Dreyfus, Hubert L. (1992) **WHAT COMPUTERS STILL CAN'T DO: A Critique of Artificial Reason**, MIT Press, London.

[Fodor and Pylyshin 1993] Fodor, Jerry A. and Xenon W. Pylyshyn (1993). *Connectionism and Cognitive Architecture*, **READINGS IN PHILOSOPHY AND COGNITIVE SCIENCE**. MIT Press, London.

[Hofstadter 1979] Hofstadter, Douglas (1979). **GODEL, ESCHER, BACH: an Eternal Golden Braid**. Harvester Press.

[LeDoux 1996] LeDoux, Joseph E. (1996). **THE EMOTIONAL BRAIN: The Mysterious Underpinnings of Emotional Life**. Weidenfeld and Nicholson, London.

[Ramachandran and Blakeslee 1998] Ramachandran, V.S. and Sandra Blakeslee (1998). **PHANTOMS IN THE BRAIN**. Harper Collins. New York.

[Turing 1950] Turing, Alan (1950). *Computing Machinery and Intelligence*, available in numerous locations online including http://www.abelard.org/turpap/turpap.htm.

[Van Gelder 1998] Van Gelder, Tim (1998). *The Dynamical Hypothesis in Computer Science,* **Behavioral and Brain Sciences** (21), pp. 615-665.

[Whitby 1997] Whitby, Blay (1997). *Why the Turing Test is AI's Biggest Blind Alley,* available online from http://www.cogs.susx.ac.uk/users/blayw/tt.html.

[Whitby 2003] Whitby, Blay (2003). *The Myth of AI Failure*, CSRP 568, available from http://cogslib.informatics.scitech.susx.ac.uk/csr_abs.php?type=csrp&num=568&id=9

[Winograd and Flores 1986] Winograd, Terry and Fernando Flores (1986). **UNDERSTANDING COMPUTERS AND COGNITION: A New Foundation for Design**. Ablex Publishing Corporation, Norwood, New Jersey, USA.

[Wittgenstein 2002] Wittgenstein, Ludwig (2002). **PHILOSOPHICAL INVESTIGATIONS.** Blackwell Publishers, Oxford.