

"The Self" -- A bogus concept

Aaron Sloman

Last updated: 24 Aug 2008; 22 Oct 2010; 14 Nov 2010; 21 Jan 2011; 10 Mar 2011

This discussion paper is at:

<http://www.cs.bham.ac.uk/research/projects/cogaff/misc/the-self.html>

A PDF version (automatically generated) which may be slightly out of date is also available

<http://www.cs.bham.ac.uk/research/projects/cogaff/misc/the-self.pdf>

Originally written in anticipation of a debate at the AAAI'08 Workshop on Meta-reasoning.

**<http://www.sis.uncc.edu/~anraja/MetaReasoning/>
July 13-14, 2008, Chicago, Illinois.**

[Needs to be formatted better.]

CONTENTS

- **Introduction: Hume on the self**
- **Galen and Peter Strawson on the self**
- **An alternative analysis: "self" and economy of expression**
- **It's not all nonsense**
- **Not all repetitions can be replaced with uses of 'self'**
- **What I am not saying**
- **I have never had a sense of self**
- **Knowing what I am going to do**
- **Knowledge of what I can and cannot do**
- **What you are depends on context**
- **The concept of an identity (Added 21 Jan 2011)**
- **What you are can change over time (Added 22 Oct 2010)**
- **Do you have a self-model?**
- **Is google part of you? Or part of your self?**
- **The biological self-nonsel distinction**
- **Dennett's notion of self as centre of narrative gravity**
- **Related discussions (a tiny unrepresentative sample)**

Introduction: Hume on the self

Over several centuries there has been much discussion of the notion of "the self"

in philosophy, psychology, theology and recently also in Artificial Intelligence. See, for example, the two wikipedia articles: [http://en.wikipedia.org/wiki/Self \(philosophy\)](http://en.wikipedia.org/wiki/Self_(philosophy))

[http://en.wikipedia.org/wiki/Self \(psychology\)](http://en.wikipedia.org/wiki/Self_(psychology))

And the superb essay on personal identity by David Hume, criticising the idea of a self as an entity:

<http://www.mnstate.edu/gracyk/courses/web%20publishing/TreatiseI.iv.vi.htm>

Galen and Peter Strawson on the self

A more recent attempt to defend the notion, by Galen Strawson is here:

<http://www.imprint.co.uk/strawson.htm>

He discusses the claim that

...the mental self is ordinarily conceived or experienced as:

- (1) a thing, in some robust sense
- (2) a mental thing, in some sense
- (3,4) a single thing that is single both synchronically considered and diachronically considered
- (5) ontically distinct from all other things
- (6) a subject of experience, a conscious feeler and thinker
- (7) an agent
- (8) a thing that has a certain character or personality

Note: I neither ordinarily conceive of nor experience any such thing.

If conditions (2) and (5) are omitted, then I do not object: there is such a thing, namely me: Aaron Sloman. There are many other such things, including you, the reader. But I (like you) am more than a mental thing. E.g. I and, I suspect also you, are things with eyes, and hands, things that occupy space, things that digest food, things that learn, that fall asleep, that will one day die, etc.

Below I contrast Galen Strawson's theory (and other similar theories) with an alternative analysis, and the closely related theory presented by Peter Strawson in 1959. My version of this theory is also closely related to (and was originally inspired by) Hume's.

An alternative analysis: "self" and economy of expression

In contrast with Galen Strawson's view, and similar views, the suggestion I offer is the simple one (with a long history, perhaps no different from Hume's in essence), namely:

Whenever anyone, X, is referring to his self, using words and phrases like "myself" "selfish" "self aware" "selfconscious", and so on, what X is actually referring to is nothing more, and nothing less, than X.

In particular it is not some "special" part of X, though what X may be saying about X can be much more complex than the form of words suggests.

People who do not understand that this is how words like "self" function in our language, regard it as legitimate to use phrases like "the self", "a self", as if they referred to some kind of thing distinct from people (or other agents), somehow forming parts of people, but not physical parts, and somehow more deeply connected with who or what the individual is -- perhaps even the source of that person's desires, hopes, fears, decisions, beliefs, etc.

(See Galen Strawson's characterisation, for example, e.g. condition (8).

Contrast Peter F. Strawson's chapter on Persons, in his 1959 book

Individuals: An essay in descriptive metaphysics, where he argues that the concept of a person, which has both physical and mental properties, is in a sense 'basic'.)

I don't know if that sort of concept is 'basic' in any sense, but whether it merits the label 'basic' or not, that is the concept that characterises what I refer to when I use the expressions: "I", "me", "my", "Aaron Sloman", namely a whole person, not any mysterious special part of a whole person.

This is an example of a common phenomenon in philosophy: some useful turn of phrase is interpreted as having a very different significance from what it actually has. The attempt to define that special significance often leads to questions, assertions, discussions that, to a first approximation, can be described as complete nonsense disguised as sense.

It's not all nonsense

I am not saying that ordinary uses of the English particle "self" (and equivalents in other languages) are nonsensical. On the contrary "self" in various contexts is perfectly meaningful as a very convenient and elegant syntactic-sugar device for use in constructs of the forms

```
..... X ..... X .....  
..... X ..... X ..... X .....  
..... X ..... X ..... X ..... X .....
```

and so on, i.e. constructs where the same referring expression, X, occurs more than once.

I.e. it allows a more compact and elegant formulation of meanings that would otherwise require the same referring expression to be used more than once.

Some examples:

John shaved himself =
John shaved John (the same John)

Mary's self control enabled her to cope with the situation =
Mary's ability to control Mary enabled Mary to cope with the situation
(all the same Mary).

Fred managed not to give himself away =
Fred prevented Fred from doing something that would reveal information
about Fred that Fred wished to conceal.

John's self-understanding prevented a mistaken decision =
John's understanding of John prevented John from taking a mistaken
decision.

and many more. In each case the use of the word 'self' or one of the related words (e.g. 'herself', 'itself', 'myself', 'themselves') does not refer to some mysterious entity that is a part of a person, but rather refers to the whole of a person (or set of persons in plural uses) whose identity is specified somewhere else in that sentence.

[*] The last sentence was somehow truncated at one stage.
It was reconstructed on 13 Nov 2010.

As shown above, in some cases more than two occurrences of the referring expression (in place of "X") get collapsed.

I could produce many more examples, but I won't, rather:

I promise to restrain myself.

(I leave it to the reader to work out why the template behind that last sentence involves at least four occurrences of 'X', possibly five.)

Not all repetitions can be replaced with uses of 'self'

The syntactic requirements for replacing one or more of the occurrences of 'X' with 'self' are fairly subtle. E.g. you can't do it with 'X is poor and X is honest'.

It is not the case that

John is poor and John is honest

can be rephrased as John something or other himself.

A sentence that can be abbreviated using 'self' requires not just properties (or unary predicates) to be used, but something like a preposition, a verb, or some other form of words expressing a spatial, temporal, causal, functional, or other relation between X and X.

I presume many, if not all, human languages share a similar syntactic device, which is likely to generate the same confusion in many cultures.

The power of syntax to generate deep clouds of smoke should never be underestimated, nor the resistance to debunking.

Idle speculation:

Perhaps in some cultures similar confusion is associated with the notion of "a sake".

How could you possibly do something for Fred's sake, if Fred hasn't got a sake?

Is there a branch of psychology called 'sake psychology', I wonder, to go alongside self-psychology? See <http://www.selfpsychology.com/>

What I am not saying

I am not trying to recommend abolition of talk using 'self' on its own or in its usual combinations 'itself', 'myself', 'selflessly', etc.

I do recommend the avoidance of certain metaphysical conclusions based on misunderstanding a useful syntactic construct.

Those metaphysical conclusions often lead the *useful* syntactic forms to be extended in *useless*, *obfuscatory* directions, often indicated by using 'self' as a common noun with the definite or indefinite article.

Such muddles are commonplace in the history of thought. Sometimes we spot the muddles easily, e.g.

'In which direction is the universe moving?'

'Is it past midnight yet at the centre of the earth?'

Sometimes detecting that a move has been made from sense to nonsense requires unusual competences, e.g. understanding the nonsensicality of this question (cf. Russell's paradox):

'How many properties possess the property of not exemplifying themselves?'

Unfortunately the analytical philosophical lobe of most brains never gets developed in a normal education.

I have never had a sense of self

As far as I know (and I don't have complete self knowledge) I have never had a sense of self.

I do, of course, acquire, store, use, and lose information about myself of very many kinds. For example I now have a sense of where I am, what I am looking at, what I can see, what my hands are doing, and what I am trying to say.

But that was just another sentence instantiating a pattern sentence with multiple occurrences of 'X', where X = "I".

And "I" (along with "me", and "my") just happens to be the most useful referring

expression for me to use to refer to Aaron Sloman. That's something I can do and you can't -- so what you hear me say, requires you, if you want to remember it and think about it, or work out its implications, to substitute "I" with "Aaron", or "him", or "you", or "that windbag", or "the author of this message" -- you have a wide range of options.

Different people have different options, depending how well they know me.

I may have more options, but they would all be very much more tedious to use than the simple "I".

I also have a sense of things I previously did, things I previously experienced, places I previously visited, things that happened to me, things other people did to me, thoughts I had, desires I had, pleasures and pains I had.

Knowing what I am going to do

It's interesting that to some extent I have information, or at least expectations, about future happenings. I now, i.e.

Sun Jul 6 10:37:54 BST 2008

know that I'll be going to the AAAI'08 conference in Chicago, as long as no disasters stop me.

Sometimes I even know what I am going to say or type before I finish. E.g. before I typed "sometimes when I type" in this sentence I knew I was going to type that sometimes when I type I know how I am going to finish the sentence I have just started.

But I don't always know how I am going to finish the sentences I start. I did not know how that long sentence would end up when I started it.

When I started writing this note, I had no intention of going on so long. (A common mistake I make.)

I am not alone in not knowing how I am going to finish utterances I start. Many people resonate to this famous quotation:

'How can I know what I think unless I hear what I say?'
(Attributed variously to E.M.Forster, Graham Wallas,
Tallulah Bankhead, and possibly others....)

Sometimes, in my case, when I hear what I say the main effect is to make me realise that that is *not* what I think, so that I need to try again to say something that expresses what I think.

The reason for this is subtle and complex: what I think is not a collection of stored sentences, but is distributed over a lot of competences and dispositions whose contents cannot easily be extracted, except by running the system. Mere inspection cannot work. But running the system in an artificial context, e.g. answering the question 'What do you think about so and so?', will not necessarily produce the same result as running the system in a real context, "with lots more of the variables bound".

The ability to do philosophy often requires the ability to short-circuit that process. I suspect that's partly a result of a lot of unconscious self-monitoring and storing various summaries of what the system does. But maybe that mechanism does not work in everyone.

Some people are much better at doing the 'short-circuiting' than others. I have found it very difficult to teach that skill: some students pick it up and some don't.

Some of the latter make all sorts of false statement about how their own minds work.

Knowledge of what I can and cannot do

Like most people I know about many things I can and cannot do. I know I can raise

my right arm while sitting here and that I cannot raise the waste paper bin at the far end of the room while sitting here.

How do I raise my arm? Just by doing it, and certainly not, as some philosophers think, by first doing something called 'willing it to go up'. I wouldn't have a clue how to do *that*!

(And would I first have to will myself to will it to go up?)

(Para added: 24 Aug 2008) My inability to move the waste paper bin remotely could change if future technology allows devices to read information about what's going on in my mind from what's going on my brain, and transfer my decisions to machines that cause objects in the vicinity to move without my touching them. If that ever happens, people may, with practice, learn to move external objects in something like the same way they move their body parts. Initially that may require some specific conscious mental process, such as saying to oneself "Bin move left". But there is no reason why such a thing should remain a necessary precursor to moving external objects any more than talking to yourself is a necessary precursor to lifting your arm. If a lot of that goes on, the already fuzzy boundaries between individuals and their environments will be even more blurred. Some of this is already happening to me (and others) in connection with search engines as discussed below.

I know I can do various things internally. I now know (after trying) that I can keep my gaze where the text is on my screen as I type and attend to the angle of slope of the desk lamp some way off to the right. (I've never done that before.)

I also know I can think of many examples of things I can and cannot do. I can recall which house I was in at 10pm last night, but not what my posture was at that time. I might be able to work out which room I was in, but I know I moved around a lot and I don't know exactly when I moved from where to where. But I can work out some of my past locations by checking broadcasting schedules and working out that I must have been in the room where the TV set is, or sitting in the kitchen where the radio is.

I can make myself choose a nonsense phrase to think, or a number between 1000 and 1000000 to think about (possibly one I've never previously named out loud). But I cannot choose to think up a valid proof of Goldbach's conjecture. I am pretty sure of that, but who knows, maybe next week I'll have a great new idea and just do it.

But I can be mistaken about where the ideas come from. More than once I have written down what I thought was a great new idea, then months or years later found it in a book I had read a long time earlier, with my pencilled comment in the margin to prove that I had read it. (Most of the ideas in this paper come from David Hume and a few of the philosophers I met while a student in Oxford in the early 1960s, but I don't recall which. Some came from Wittgenstein's *Philosophical Investigations* I think. Several come from hearing Gilbert Ryle and John Austin lecture, or sitting in their seminars around that time.)

Most people have no idea where or from whom or what they first learnt the vast majority of the words they use, or the concepts they use, or the facts they have learnt. In general the information is not worth storing because it is of no use. Academics concerned about plagiarism have to learn to remember unnatural things.

How do I know some things about myself, and why don't I know others? Those are questions to be solved by future collaborations between AI, psychology, neuroscience, and even philosophy, -- though there are fragments of answers, e.g. concerning which bits of brains make a difference to different kinds of mental competence.

However, not only is it the case that we can be mistaken, there are also very serious pathologies in which the inability to do all these things normally can make people highly dysfunctional, even dangerous, e.g. schizophrenia -- a deep and multi-faceted disorder, in which many of the things I've been talking about stop working properly.

All that is just a long-winded way of saying what I am not saying when I say I don't have a sense of self.

What you are depends on context

A person can take on different competences, preferences, likes, dislikes, reactions, ways of seeing, ways of acting, ways of speaking, ways of reacting to people, in different contexts.

(This is not usually done deliberately, or even noticed by the individual.)

This is probably true of many intelligent animals: a lioness needs very different collections of switched on competences and readily available information (a) when lolling in the sun with her cubs, (b) when a strange male lion approaches as a potential mate, (c) when she is stalking her prey, (d) while she is actively chasing a terrified deer, and (e) when she has caught and killed her prey, etc.

A person who is a charming and entertaining father playing with his children can turn into a maniac with road rage while driving his car, can become an overbearing, arrogant manager dealing with underlings at the office, and turn into a pathetic, fawning, nervous wimp when talking to his superiors.

Of course, he need not be aware of any of this: often our acquaintances know more about our state of mind than we do (a common theme in novels and plays -- for instance about infatuation or jealousy).

We can compare these different states with different states of a complex software system that can load different control parameters and rules to control its behaviour in different contexts, e.g.

- in the early stages of booting up,
 - when an intruder has been detected,
 - when maintenance engineers are looking for bugs,
 - when file space is running low and permissions and quotas have to be adjusted,
 - when connected or disconnected to a potentially dangerous network,
- etc.

This is why, the impressionistic diagrams representing the H-CogAff architecture have a box labelled "Personae'" referring to alternative global control states between which an individual can switch according to requirements of the context:
<http://www.cs.bham.ac.uk/~axs/fig/your.mind.jpg>

Perhaps I should have used the label "selves" or "personalities"???

The switching needs to be to some extent automated -- e.g. it could be one of the functions of the reactive 'Alarm' mechanism. If it weren't *automated*, the system currently in control might not wish to relinquish control. Or it might take too long.

As suggested in the 1996 message listed below, it is possible that dysfunctional versions of the mechanisms involved in switching what I have called "Personae" could account for at least some examples of so-called 'Multiple personality disorder' (MPD).

Hypnotism may depend on the fact that the switching can, to some extent, be triggered from outside. But there need not be a sharp boundary between hypnotism and other social influences, e.g. in teaching, preaching, inspiring, threatening, tempting, advertising, application of peer pressure, etc.

The concept of an identity (Added 21 Jan 2011)

I am grateful to [Yasemin Erden](#) for reminding me that some ordinary uses of the

words "identity" "identities" overlap with references to what I have called Personae, though some notions of identity go beyond the notion of the current global control-state of a mind.

For example, a con-man can adopt different identities as well as different personalities, in different contexts. A fake identity might be determined by a false passport without any fake personality being involved. This is a very old and useful idea, which has acquired a wider range of uses since the development of computer games and various types of internet based interaction, or even some board games ("I'll be the butcher this time.").

For example, you can adopt different identities in different email lists, depending on what you wish to reveal about yourself (sic!) to other members.

What you are can change over time (Added 22 Oct 2010)

Many biological organisms (members of "precocial" species) are born or hatched with fixed information processing mechanisms, perceptual capabilities, motor capabilities, motives, drives, and behavioural competences. Alternatively they may change slightly through adaptation to the environment in a process that sets parameters.

[*]In the next paragraph I originally typed "precocial" instead of "altricial".

Fixed 14 Nov 2010.

Other organisms (members of what biologists call "altricial species")[*] can change many aspects of their information processing architecture and the resulting competences, along with forms of representation, ontologies, knowledge, values, preferences, goals, attitudes, ideals, etc., during learning and development.

Humans are capable of going on developing in such ways throughout their life span. How such changes happen and the complex interplay between genetic and environmental influences and the growing control by the individual of how that should happen, are very complex processes based on mechanism that are barely understood at all, and are certainly lacking in current so-called intelligent robots. (Future robots may be different.)

Jackie Chappell (School of Biological sciences) and I wrote an invited journal article on this topic, combining biological, AI and philosophical ideas. It is freely available online:

Jackie Chappell and Aaron Sloman,
Natural and artificial meta-configured altricial
information-processing systems,
International Journal of Unconventional Computing, 3, 3, 2007, pp. 211--239,
<http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609>

Do you have a self-model?

Some people who get involved in discussing 'the self' are actually claiming that people have models of themselves. There's nothing wrong about that claim if having a model of yourself is something like having a theory about what sort of thing you are, what sort of person you are, how you work, what you know, what you can and cannot do, what you like, hope for, dislike, etc. and perhaps many other things.

In that sense I have a model of myself too, though it is not a unitary *thing*, and it is not static -- I change my mind about myself over time, and sometimes I learn new things about myself, e.g. from honest acquaintances (especially my wife, who has several times had to draw my attention to facets of myself that I had not noticed, especially facets involving behaviour towards others in discussions).

But the fact that we can say that Fred has a model of himself is not an exception to the claim about syntactic sugar above, for it amounts to nothing more or less than the fact that we can say that Fred has a model of Fred.

And that is just a piece of jargon for saying that Fred has knowledge, beliefs, conjectures, expectations, hopes, fears, memories, and other kinds of information about Fred -- not all of it correct.

Some people want to relate this notion of having a self-model to the notorious phrase made famous by Tom Nagel in his 'What is it like to be a bat?' ([Available here](#))

I think the notion of 'what it is like to be something' is also a bit of bogus syntactic flummery which is about as meaningful when taken out of sensible contexts as 'what time is it at a place'. (What time is it now at the centre of the earth?)

As a deflationary exercise, I once wrote a paper on what it is like to be a rock. It annoyed some readers and amused others:

<http://www.cs.bham.ac.uk/research/projects/cogaff/misc/rock/>

Is google part of you? Or part of your self?

Many people seem to have discovered a new way of dealing with the 'tip-of-the-tongue' phenomenon: you know you know a name or word or phrase but just cannot remember it, no matter how hard you try. Later, when you are doing something completely unrelated, you suddenly remember it.

Instead of giving up and waiting for the brain to do its stuff in its own time, which used to be the only alternative (apart, perhaps, from attempting systematic searches, e.g. using the ordering of words in the alphabet), some people (including me) have discovered that they can go to google, and type in some carefully chosen related words and phrases, so that the desired item turns up near the top of google's output. As the difficulty in recalling things on demand grows with age I have been relying on this technique more and more.

To that extent, google has now become part of me, even though it is not always available. Not all of what I know is always available either.

The biological self-nonsel self distinction

None of what I have written is a criticism of the talk by biologists about organisms using mechanisms to distinguish self from non-self. This is just a manner of speaking about real functions of things like the immune system and other mechanisms that have to detect and deal with abnormalities, waste-products, threats of various kinds, etc.

It was Catriona Kennedy who first drew my attention to that usage, and at first I mistook it for another example of the philosophical confusions discussed above.

Dennett's notion of self as centre of narrative gravity (DRAFT: 24 Aug 2008)

See <http://cogprints.org/266/0/selfctr.htm>

We have many ways of thinking about what we are or have been or will be doing or experiencing. Dennett's notion offers an analogy, partly based on a piece of verbal dexterity, between a character in a story and the centre of gravity (he

probably really meant the centre of mass) of a physical object.

He then assumes that an object's centre of gravity (or mass) does not really exist: it is just a fictional object. This is an error, since the centre of gravity is no more fictional than the longest dimension of an object, or the point on its surface with maximum curvature, or the highest point of a mountain range. The centre of mass (or gravity) of an object at any time is a location in space (often relative to frame fixed in the object, if it is a rigid object, though other frames can be used) mathematically defined by its relation to the distribution of mass throughout the object.

The intended argument seems to be that, where instead of a story we have a real history of some individual, that story has a "centre of narrative gravity" (CONG) which is related to all the parts of the history in something like the way in which the centre of gravity of an object is related to the parts of the object. Then just as the object's centre of gravity is alleged to be just a useful fiction, so, suggests Dennett, is "the self" associated with an individual, analysed as the CONG associated with that individual, just a useful fiction.

I think this is completely unhelpful because there really is a non-fictional entity associated with the history of Daniel Dennett, namely Daniel Dennett, and he, that particular human being, has all the features required to be himself!

There is another objection to the analogy: whereas a physical object has a unique centre of gravity, I, like most people, have a host of different locations simultaneously because I am embedded in different spaces that may have little to do with one another. Some of these are similar to a centre of gravity: E.g. if I am in a train I have a location in the train, and if I move to another compartment I have a different location in the train. At the same time I have a rapidly changing location along the railway line, and I may be moving along it in one direction and simultaneously moving in the opposite direction on the train.

I also have locations in and in relation to many other structures that are important, some of them quite abstract (e.g. social networks, roles in organisations, etc.), and the variety of such locations has exploded in the last 30 years because of the development of computers. (I haven't even mentioned computer games using virtual worlds so far.) I may also have a place in a novel I am reading, and since I usually have several downloaded research papers open at once on my computer I usually have locations in all of them, remembered for me by the software tool that I use to read them (e.g. xpdf remembers the page I am on in reading a PDF file, if I leave it open). Likewise in things I am working on -- papers, email messages, book chapters or whatever: I can have locations in all of them at the same time.

So IF there's any use for the notion of a centre of narrative gravity then I, for one, have many of them. I expect you have also.

I don't think I have a substantive disagreement with Dennett (he too quotes David Hume with approval): only a difference of style, and a different view about how to help self-theorists discover what is right and what is wrong in their theories.

Related discussions (a tiny unrepresentative sample)

- McCarthy on 'Making Robots Conscious of their Mental States'
<http://www-formal.stanford.edu/jmc/consciousness.html>

- Minsky (1968) Matter Mind and Models
- Look for mentions of 'self' in Minsky's *The Emotion Machine* (linked on his web page): <http://web.media.mit.edu/~minsky/> (Especially Chapter 9 "The Self".)
- My notes for a DARPA workshop in 2004 on self-aware machines
- 1996 Posting on self control to the Psyche-D discussion group
- Four concepts of free will, two of them garbage.
- I have a short tutorial on conceptual analysis in this paper on Varieties of atheism.
- Presentation at Bielefeld (Oct 2007) on Why robot designers need to be philosophers and vice versa.
- "New Bodies for Sick Persons" (1971)
About curing cancer by making a copy of your body without the cancer cells, and making sure the original doesn't wake up.
- Gilbert Ryle *The Concept of Mind* (1949) is well worth reading. Unfortunately it is not yet online, as far as I know.
(If anyone tells you Ryle is a behaviourist point at the chapter on imagination.)
- Most programming languages give running programs very little access to their current states and processes.
Exceptions include Lisp and Pop-11
- Chapter 9 ('Being No One') of Shimon Edelman's new book *Computing The Mind: How the Mind Really Works* (OUP) 2008.
- Galen Strawson's paper:
<http://www.imprint.co.uk/strawson.htm>
Alas, much philosophical writing on this and related topics could benefit from adopting the design stance, as summarised here.

Initially created: 5 Jul 2008

Modified: 21 Jan 2011

Maintained by Aaron Sloman
School of Computer Science
The University of Birmingham