Representational Structures in Consciousness

§1 Description, not Explanation

This essay tries to elucidate structural elements of consciousness by drawing on phenomenological descriptions of consciousness and employing representationalist models of consciousness. The aim is not to explain the genesis of self-awareness, but to outline different aspects of its structure.

Given the vast literature on consciousness some key terms have been given ever so slightly different readings (e.g. ‘reflexion’, ‘attention’, ‘self’, ‘immediacy’ etc.). Some controversies seem more to be a setting apart of theories on certain reading of key terms than undissolvable contrasts in outlining structures of consciousness. Critics of some type of account sometimes like to convict a theory quite similar to theirs on account of employing some supposedly inappropriate terms. Some terms are preferred here, and some are given an explicit reading or use. The essay aims at a structural *description* of consciousness. Although the description is developed in contrast to Higher Order Theories or appeals to ‘reflexion’, the descriptions provided, despite expressing a *Same Order Theory*, may not be so different from those of some version of some theories in the field od Higher Order Theories. All this illustrates the difficulties of establishing a proper vocabulary for the study of consciousness. Like in other parts of science our everyday linguistic conceptions of middle-sized objects may mislead us.

§2 Structural Differences

We may start with a sentence expressing a state of affairs

(1) The pen is black.

(2) I believe the pen is black.

(2) expresses a mental state of consciously or sub-consciously believing (1), i.e. believing that the state of affairs expressed by (1) obtains. We can use (2) to talk about conscious or sub-conscious beliefs. To capture the contrast between conscious and sub-conscious beliefs

we use (2) here as attributing a sub-conscious belief. In contrast to (2) one may express a conscious belief by

(3') I consciously believe the pen is black.

The tradition of theory of consciousness points out correctly that in consciousness I am *aware* of what *I* am doing, i.e. of me as the thinking agent, although typically the focus is on the content state of affairs not on my thinking. This structure, captured in the thesis that all human consciousness is conscious of the thinking agent (i.e. is self-aware in a sense to be further elucidated) is better captured in

(3) I am conscious of me believing the pen is black.

The agent is non-focussed (non-explicit) content in (3). ('implicit' may be misleading because of the notion of 'implicit knowledge', which is not conscious at all.)

(4) I am conscious of ME believing the pen is black.

(4) brings the thinking agent *into focus*. This focussing is *not* the introduction of higher order thought. Kant captures this as the 'I think' that *can* accompany all thought (i.e. one *can* focus on oneself), but it need not, as typically, (3), consciousness is not *focussed* on its agent.

(5) I believe I believe the pen is black.

(5) shows the structure of higher order belief. A conscious higher order thought is expressed by

(6) I am conscious of me believing I believe the pen is black.

A higher order conscious state can also have being conscious as part of its content, expressed by

(7) I am conscious of me believing I am conscious of me believing the pen is black.

All these states expressed by (2) – (7) are different. Thus

(mere) representation ≠consciousness

explicit self-representation ≠ consciousness ('explicit' in the sense of 'focussed', s.a.)

higher order thought ≠consciousness

A representation can *represent* the thinking agent without being conscious, e.g. (2). A representation can be higher order without being conscious, e.g. (5). A representation can be conscious without being focussed on the thinking agent, e.g. (3) vs. (4).

Most conscious acts are straightforward (i.e. focussed) on the world, not the act or agent if the act. Consciousness is thus mostly transparent. A shift of focus on the agent of the conscious act is not a reflexive act (not generating a state of higher order). (Husserl at times speaks of an aspect of consciousness as 'conscious but not recognized'. Recognition in this sense is not reflexion.) 'Higher order' respectively 'reflexion' is a notion of structural embedding or iteration. Shift of focus is a different operation.

Consciousness is monolithic in the sense that one can have and be conscious of higher order states (embedding states of the same or different kind), but one cannot have *a second* consciousness within one's consciousness. So

(8) I am conscious of me believing I am conscious of me believing the pen is black.

expresses a conscious state which is conscious of a belief *about* consciousness. Consciousness is represented in the belief I am conscious of. But

(9*) I am conscious of me being conscious of me believing the pen is black.

is – given the view put forth here – a misrepresentation of a higher order belief. (9*) does not occur as an act of thinking.

In

(4) I am conscious of ME believing the pen is black.

the "I am conscious of" represents the occurrent process of thought with the thinking agent. This agent is represented as "I" since it is to be identified with the "ME" (or "me" in (3)) in the content of consciousness. The 'me/ME' has a phenomenal quality as representing the agent of thinking. (In nowadays parlance one might say 'the what it is like of what it is like'.) In it the agent of consciousness is *conscious* agent. (In Sartre's notation the conscious agent is conscious "~~of~~ itself", i.e. immediately, not in an act of reflexion [i.e. a higher order state, or in Sartre's terminology 'improper reflexivity'].) The expression "I am conscious *of* me ___" articulates that the thinking agent (the subject) is aware of itself as involved in this or that act

of the thinking agent; it does not articulate itself as just another object – not as "I am conscious *that* I ____". It articulates itself as relating to itself as the agent of those acts.

Every consciousness has self-access be it focussed or not. There is no 'ego-less' consciousness. Talk in this way is misleading way of expressing the difference between focus on content, (3), and focus on thinker, (4). In straightforward conscious states like (3) the subject is not in focus. In a limited sense one can thus say that 'the given is subject-less', a formulation which stresses the perceptual objective contact to reality, but is misleading in the philosophy of mind.

The activity as activity (something being non-static) cannot be *exhausted* by a representation (something static). Self-access possesses the quality of a lived experiential process.

(Again in Sartre's terminology: the agent is always 'exstatic'. Fichte tried to capture this difficulty with the image of the Ego being 'an activity with an inset eye'. Natorp declines such images and puts the Ego completely beyond representational content, and thus beyond being represented, apart from the minimal representation "the x that does the representing". This does not seem to match phenomenology as we experience ourselves *as* agents of thoughts. Natorp does not decline that in every conscious act the relation to the Ego is phenomenally given, he declines that this relation can be made the sole *object* of the occurrent thought. He emphasizes in this way the *unity* of the conscious act as not consisting of a consciousness of the object and a *second* consciousness of the conscious act, cf. (9*).)

There is no way around this difficulty of the unity of self-access and activity by appealing to reflexion, as reflexion involves the distinction between reflecting (active) and reflected conscious agent (object of the reflexion). This may be a unique difficulty characterizing the thinking agent. Nonetheless one must try to work out some form of expressing the self-access, because without some way of expressing it, the thinking agent becomes ineffable, and the theory of consciousness shorter than it already is. This self-access and -awareness is a moment/part of the conscious act. As such it can be made the object of a description and a theory. One cannot, however, make it the sole object of a conscious act, inasmuch as it is always accompanying the consciousness of something. Thinking about self-access – presumably in language – makes self-access the object of thought, but in this thought the occurrent self-access of the agent of this thought is not objectified (apart from being one of the species the thought is about), but exstatic (alive). Even if the acting Ego is always something exstatic (and thus in a sense is always escaping being completely objectified) this

does not preclude developing a theory of this process and its structure. Compare: a theory of time, which is a theory of a process – at least in the 'tensed' view – which itself takes time to express and comprehend, but nonetheless covers time and temporal entities (like expressing the theory itself).

Self-consciousness being *sui generis* (i.e. not a subject-object relation, not a reflexive relation, not an identification with an already somehow known self or 'I/Ego') can hardly be expressed in our common language of intentional states and propositional judgements. Attempts to point to its *sui generis* status have to use otherwise aporetic phrases like "immediate self-awareness" (sounding like an awareness *of* a self, which nonetheless should be immediate), or "non-relational/non-thetic self-consciousness" (which nonetheless is awareness the thinking agent has of himself). Self-awareness has features that highlight that notwithstanding its non-relational and immediate quality moments of it (like awareness, agency, subjectivity) can be distinguished and have to be present in an unbreakable unity.

One may say that immediate self-awareness is 'ego-less' highlighting that this awareness should not be modelled on the subject-object scheme, for all the traditional reasons based on different regress arguments. In that way of talking any object 'ego' is transcendent with respect to immediate self-awareness. Nonetheless does immediate self-awareness include knowing *of* the thinking agent in the way that the idea that somebody else might have these thoughts does never arise. It is preferable, therefore, to speak of the thinking agent as 'Ego', and of immediate self-awareness as a state in which the Ego is immediately (not mediated as an 'object') present to itself. This very state seems *sui generis* as it cannot be modelled on the form of ordinary intentional acts – leaving moods, to be another problem, to the side for the moment. Its extraordinary nature lies at the heart of many convolutions in theories of self-consciousness and many of their *cul-de-sacs*. It being *sui generis* stands in the way of explaining it within a theory of consciousness. It seems to be beyond theory. One can, however, describe representational forms of consciousness, like (3) or (4), in which occur some special representations and representational combinations, like "I am conscious of me ___", sometimes – only slightly sarcastic – conceived as an 'I-symbol' the tokening of which results in the presence of immediate self-awareness. Talking of an 'I-symbol' also elucidates that self-consciousness does not create itself. The tokening of the 'I-symbol' gives rise to a conscious state, but the symbol does not do the tokening. In a cognitive system (a person) certain symbols are tokened into complex representations, some of which are the representations underlying/giving rise to consciousness.

§3 The Ego

'Ego' shall refer to the agent of thinking. 'thinking' always means conscious thought in distinction to mentation (mental events which are not conscious). The thinking agent in its role as thinking agent has no specific gender; pronouns can be used at will. The Ego should be distinguished from the occurrence of cognitive agency on a sub-conscious level (a 'functional Ego') and a narrative self-representation of a thinking agent's biography and self-understanding: the 'Self'.

A tokening of a certain, specific representation results in a conscious state. This representation is a representation of the thinking agent in *this very act* of thinking (and tokening conscious representations). This is the occurrence (a process) of representations which cover the occurrence itself. The thinking agent as agent is always not just represented but active in thinking and representing. Being active does not exclude being represented as active, but this representation *of* the agent is not the agent. (One of the traditional problems of having a theory of the 'transcendental ego'.)

Talking of a thinking 'agent' and conscious 'acts' does, of course, not mean that conscious acts are actions. Actions result from conscious acts like beliefs and volitions, so these (on pains of a regress) cannot be actions. Talking of 'acts' stresses that consciousness develops in (inner) time as flow and that we are present to ourselves as the agents of our conscious ongoings, even if some content is pressed upon us (e.g. in perceptions) by our situation. Even if some painful event happens to us, we experience ourselves as the ones who 'do feel the pain'. Part of the Cartesian evidence is that 'I am thinking' (i.e. both being and being actively conscious), not that thoughts merely happen to me.

In a conscious act the Ego is thinking agent and part of the content. We know the Ego in its activity, not its ontological essence or metaphysical nature. Thus – as Kant stressed – awareness of ourselves as cognitive agents is compatible with our ignorance about the ontological nature of ourselves – say as brains or souls.

The topic is consciousness or awareness in humans. One part of the theory is the traditional claim that this always involves some form of awareness of the thinking agent (the Ego). In that sense every human consciousness is self-consciousness/self-awareness. (A thesis going back at least to Aristotle's *De Anima,* and an explicit critique of reflexion or higher order theories of consciousness developed in the tradition from Fichte to Sartre.) This, however, should not – as unfortunately in parts of the 'philosophy of mind' – be confused with

consciousness of the self, ‘self’ understood as the narrative, biographical self-representation of a thinker/person. The ‘self’ in this sense (and so is ‘self’ employed here) is not the Ego. The Ego is what often has been called ‘the I’. It will have a correlate in non-conscious mental acts, which may be called the ‘functional Ego’; that correlate plays no great role in our considerations here.

(In a reversal of terms, but in stressing the same point, Sartre speaks of the ‘transcendence of the Ego’ [i.e. the self] with respect to consciousness, which always is immediately [i.e. without reflection] aware of itself. All these terms have been abused and abused in different theories of self-consciousness. Hopefully my use of “Ego” will become clearer in the context developed here. I do not think a completely new term can be helpful because of the intimate relation between our personal use of “I” and self-consciousness. The Latin expression ”Ego” indicates its role as a theoretical term in a philosophical elucidation of self-awareness. It also helps to avoid confusions with talk of the embodied person or the biographical self-narrative.)

I may say that some experience happened to me or that I underwent an experience. I was experientially confronted by something. I do not say that some thought (especially in inner speech) happened to me. Sometimes a thought ‘springs to mind’ or one ‘has an idea’, but these are not the standard cases. In thinking (in inner speech) I graft thoughts, ponder ideas and questions, and reflect on these thoughts or thought in general (including experiential encounters) In all this I experience myself as a thinking agent. My conscious life not just happens to me. I am not an observer of some conscious life, I am engaged and ‘participating’ in my conscious life.

The agent of my thought as conscious is not distinguished from myself, i.e. there is no anonymous agent and a personal Ego, but the person experiences herself as the agent of her thoughts. This agent is not the narrative (biographical unity) in which a person understands herself (the ‘self’ as object), because a thinking agent is not a narrative (a narrative is a static object). Nonetheless we are conscious of ourselves also in the sense of conscious of our selves, as we always understand the thinking agent as a phase of a larger whole we understand as our self.

The unity of conscious content became unified as contents of one Ego. We understand ourselves in correlation to the whole of our conscious life. We generate a self(-description) out of our lived experience. In this way unity of self and unity of content are correlational. Nonetheless the activity of the thinking agent is the activity which establishes the correlation.

The relation is not symmetric, only in our understanding do both sides depend in each other. In the broader metaphysical picture the thinking agent depends on the environment it is embedded in (i.e. the referents of some of the representations in consciousness). The metaphysical conditions the conscious agent depends on need not be unified into the same whole the content of its thoughts are unified into.

§4 Representations and Content

Conscious self-access is of one type: 'I am conscious of me ___' does not come in different qualities. Qualities and types occur in the states one is conscious of: seeing, hearing, seeing something red, something blue. The content of consciousness harbours a manifold of qualities and represented objects apprehended by states of different types (including believing, desiring etc.), whereas the aspect of being conscious in these acts is always the same.

In

(3) I am conscious of me believing the pen is black.

the part "the pen is black" expresses the content of the belief, and thus the content of the consciousness. This content is present in a belief-type state. The whole state 'believing the pen is black' is content of the consciousness. Also content ('agent content') of the consciousness is 'me believing the pen is black'. We have thus three types of content:

a) objective content (the state of affairs represented by a sentence or a pictural representation)
b) state-type content (the mode in which the content is present)
c) agent content (representing the thinking agent)

Representational structures are also important because of their functional role. Two sentential representations may share their referential/informational content, but differ in the role they play in the mental life of a thinking agent. This difference has been the cornerstone both of theories of oblique contexts as well as theories of essential indexicals or 'indicators'. A temporal expression may refer to the same time as the indicator "now", but the functional role of beliefs involving the different expressions may vary widely. Similar remarks apply to the subject's use of "I" and the use of a representation that objectively identifies the thinking agent for an audience. A theory of the mind interacting with others and reality thus must

involve a theory of indicators and the representational format of self-access. The functional role of a representation in consciousness does not reduce to its referential content. Therefore, indicators like “I” and “now” are ineliminable for such a theory.

Within some version of the Representational Theory of Mind – employed here – the attitudes are taken to be sentential/propositional attitudes. The content of (conscious) mental states can be expressed by sentences of some internal or external language. These sentences as employed in a situation of usage have referential content (computed by anchoring indexical expressions to appropriate entities). As representations these sentences also have a mode of representing their referential content. The same referential and even the same semantic content can be represented by different sentences. Understanding attitudes and content as involving such (tokenings of) sentences accounts for the hyper-intensionality of attitudes. The way or mode of representation often results in a different functional role of the individual mental event. Informative identity statements have the same referential content as uninformative, but consist in a mode of representation with a unique functional role. E.g. knowing ‘Monday is Monday’ and knowing ‘Now is Monday’ differ not in their referential content, but in its mode of representation, which accounts for the different functional roles corresponding beliefs may have. A Representational Theory of Mind, thus, allows for indicators/indexicals in representations of mental content, without denying that there is an objective referential content. And objective referential content does not exclude a perspectival mode of representation from a subjective point of view. Accounting for the representational structure and status of self-awareness does not require the introduction of special objects of self-ascription in self-awareness (e.g. special properties/predicates vs. propositions/sentences) nor special individualized kinds of the attitudes (like ‘believing-of-oneself that ___’ vs. ‘believing that ___’) beyond the distinction between referential content and representational mode. Representational content is referential content represented in some way/mode/character.

The pen, in (1), is an enduring object. A perception of the pen is a fleeting experience. A constant perception unites a series of successive or overlapping perceptions into a longer period of a constant perception of the pen as the objective ground of the series. In the series or movie an enduring perception of the pen corresponds to the enduring pen. In a similar way the Ego presents the thinking agent. Conscious acts are united (instantly) diachronically by retention and protention, and synchronically by a form of immediate conjunctive equivalence of the form “$I(\varphi \,\&\, \psi) \equiv I\varphi \,\&\, I\psi$” into the stream of consciousness. We experience a *single*

(not several co-occurring) and *lasting* Ego. This experienced lasting Ego does not contradict the fact that each conscious act has an Ego, no less than the enduring movie representation of the pen contradicts the presence of a pen picture in each movie frame. The subjective experience of a lasting subject of consciousness corresponds to the lasting thinking agent as the underlying object. The Ego is the self-presentation of this objective ground, be it the brain, the soul, or whatever.

For perceptual mental states awareness of the state is constitutive. In case of beliefs and other mental states that can be had without them being conscious the more complex structure (3) brings out the difference between a conscious belief and an unconscious one, like (2). Some structure of self-representation has to be involved – or better: the belief has to be involved in this structure – to switch from an unconscious belief to a conscious belief. Whereas we can have non conscious beliefs, mental states involving perceptual/phenomenal concepts (like: 'seeing', 'hearing' …) are *always* conscious. The concept 'black' involved in the belief

(2) I believe the pen is black.

is – if not different – though only partially (referentially) identical to the phenomenal concept in

(10') I see a black pen.

Phenomenal qualities occur only in conscious states. An equivalent form of (10') is

(10) I am conscious of me seeing a black pen.

while (2) and (3) are not equivalent. An act of seeing does not become conscious by being the object of a higher order state or by *becoming* embedded into a more complex structure involving self-representation. It exists only in those structures. There are different constraints on the mental use of concepts which are not phenomenal and those which are. The presence of an 'I'-symbol may be part of an elucidation of the presence of self-awareness (as ever-present aspect of consciousness in humans), still there are other constraints with respect to phenomenal qualities in consciousness, constraints of a type and function which – although we have no idea how – may have precursors in those animals which have awareness (without self-awareness).

§5 Reflection

In reflection one act is the object of another act, cf. (5). This reflexive act can be reflected upon. The object of reflection then is a more complex act represented by a more complex representation involving embedded cognitive state (i.e. embedded cognitive operators). Iterated higher order reflection thus *represents* ever more complex states of embedding. From a *procedural* perspective there are always *only two* levels: the ongoing mental process (be it one of thinking about thinking) and its represented object (be it acts of iterated thinking about thinking).

Reflexive or higher order states are pervasive in our mental life (e.g. in the dynamics of belief acquisition, update and revision, or in the following of conventions). Consciousness, however, cannot (for the reasons present in the philosophical tradition starting with Fichte) be explained or modelled by a Higher Order Theory – and it *should not* be so explained in the light of good theories of conventions and belief dynamics, as the higher order states involved in the background (implicit) reasoning are *not* conscious and would overload conscious life if any higher order state was conscious. Not any higher order state, thus, must be conscious. If the mere presence of higher order states yielded consciousness, even artifacts which we do not consider to be conscious (as some IT-systems with self-monitoring representations of some of their states) had to be conscious.

If a Higher Order Theory now liked to make a distinction between those higher order states and those leading to consciousness, the theory shifts to those distinguishing features as defining consciousness, instead of placing the crucial feature in the higher order structure.

Anyway, the higher order state supposedly presupposes the existence of the first order mental state, and thus follows it *in time*. This delay may be one problem to ascribe causal powers to conscious states. A further problem in this vain rests in the higher order state not changing the given first order state, so that the causal powers/functional roles associated with consciousness cannot rest in that state, but only in the new state, i.e. the higher order state, which *itself* is not conscious. Causal efficacy of consciousness gets lost or cannot be explained.

It is – at best – misleading to speak of a mental state/act ‘becoming conscious’. Non conscious mental states and conscious mental states are structurally distinct. What ‘becomes conscious’ is the content (state of affairs) that is represented in the non conscious mental

state. The content is conscious content in a conscious mental state, a representation token of that content is present in the conscious state. The non conscious state is not present in the conscious state, and it does not undergo some transformation to become conscious. Why some *content* enters consciousness is a epistemological/psychological important question. It points to some procedures of tending to epistemic issues or pressing decisions to be made by the cognitive agent. The mental operations and the (non conscious) self-monitoring of the cognitive system may give rise to conscious states, and if there was an explanation to be had for why some content becomes conscious, it might be found here. The structure of conscious states does not *explain* – as some Higher Order Theories pretend – why some content is conscious, it *describes* what is involved in consciousness.

A non higher order representational account of consciousness resembles a Higher Order Theory by emphasizing the role of embedded (self-)representation for the occurrence of consciousness. Embedding (e.g. in the scope of operators), however, need not be higher order in the sense of a state being the object of another state (i.e. in the sense of reflexion of Higher Order Theories). A representationalist account that relies on the – traditional – idea of immediate self-access or self-presenting representation can be taken as a 'First Order Theory' of consciousness.

§6 A Fixed Point of Representation

From a formal perspective we can consider consciousness to be a *fixed point* of representation. We can start with the relation

(IR) x represents y as representing z as φ

in which some object represents the representational capacity of some object with respect to another object and its properties. Formal syntax allows self-application of predicates by diagonalization of predicates, which yields fixed points for predicates (i.e. objects x such that $x = f(x)$). We know from meta-logic that there provably are such fixed points for Σ_1-relations (as self-referential sentences "s = φ[s]" with "[s]" being the representation of the sentence s) by employing diagonal functions and some representation scheme (like Gödel-numbering).

(IR) allows setting some of its arguments to be identical. Thus, we may get

(FR) x represents x as representing z as φ

Informally this may be expressed as

(FRI) x represents itself as representing z as φ

And this is structurally as some instance of (3) or (4)

(FR3) I am conscious of me as representing z as φ

Self-representation, self-reference and being a fixed point do not *explain* consciousness – as, obviously, not any old fixed point yields consciousness – but the example of fixed points makes self-representation understandable and demystifies it. The self-representation occurs in the very act of representation itself. Only in this way can the causal powers of this act involve causal efficacy of being conscious.

Corresponding to (IR) we may think of a faculty of representation, and if this faculty turns on the representative nature of some of its representations employed, it might harbour a fixed point of this faculty: a representation representing itself as representing. The viability and finiteness of fixed points in meta-logic (concerning finite syntactic structures) fends off the idea that this needs to involve infinite representations. "s = φ[s]" means s = φ[s] = φ[φ[s]] etc. but we need not have all these (longer) representations, as the fixed point itself suffices. The fixed point (like the well-known Gödel-sentence) is a *finite* structure.