The Parting of the Ways: Political Agency between Rational Subjectivity and Phenomenal Selfhood

"We must learn to dissociate subjectivity from selfhood and realize that if, as Sellars put it, inferring is an *act* – the distillation of the subjectivity of reason – then reason itself enjoins the destitution of selfhood." – Ray Brassier, *The View From Nowhere*

1. Introduction: Two Kinds of Subjectivity

The above quote is taken from a paper by Ray Brassier, in which, while championing the important insight of Thomas Metzinger's theory of **phenomenal self-consciousness** – that there are no such things as **selves** – he nevertheless criticises Metzinger for failing to distinguish this from what Wilfrid Sellars, following Kant and Hegel, would call **rational self-consciousness –** or that there are nevertheless such things as **subjects**. In essence, he takes Metzinger's everlasting contribution to be finally allowing us clearly distinguish between **phenomenal selfhood** and **rational subjectivity**. The aim of the current paper is to explore this distinction further, by showing how Metzinger's theory of phenomenal consciousness can be integrated into a broadly Sellarsian theory of rational consciousness, along with some of the consequences of doing so. This is made possible by the fact that both theories are forms of **functionalism**: they aim to describe how a **causal system** would have to be configured in order to be conscious, independently of its material substrate. This means that they apply both to the *biological* forms of consciousness that have evolved on this planet, and to the *technological* forms of consciousness that we may one day construct.

The divergence between Metzinger and Sellars' approaches comes out of a difference in emphasis regarding the two aspects of the traditional notion of subjectivity: **awareness** and **agency**. Metzinger provides a functional description of *consciousness* as a form of **information processing**, supplementing it with a functional description of *self-consciousness* that, in presenting it as a *modification* of this sort of information processing, turns the 'self' into so much informational mist, letting its substance evaporate under our theoretical gaze. It does this by understanding the way a system becomes aware of both its 'world' and its 'self' in terms of the way it **represents** them. I will show that a Sellarsian functional description of consciousness can be built on top of this account of information processing, but that this involves understanding the way a system becomes

an agent capable of undertaking **theoretical commitments** about its world, and **practical commitments** about how it will act within it, in terms of its **responsibility** for them. This lets us see the subject as the stubborn **locus** of responsibility that refuses to sublimate along with the self. It amounts to a form of **normative mineness** that will nevertheless be shown to have an intricate relationship with **experiential mineness**, the consequences of which can only be sketched here.

2. Metzinger: The Dissolution of the Self

Before we can present Metzinger's theory, it is first necessary to say something about explanation itself. It is entirely possible to hold that nature is composed of a single **metaphysical plane**, that there is a single **mode of individuation** of entities upon this plane, and only one **form of causation** between these entities, while nevertheless admitting that there are many **forms of explanation** that can be applied to all of this. **Ontological univocity** is compatible with **explanatory equivocity**. [unpack]

We won't say anything more about non-causal forms of explanation. [brief examples] The salient fact for our present enterprise is that we describe causal systems by applying **explanatory schemas** to them that facilitate the development of **predictions** about the way that they would behave under various possible conditions. These schemas provide us with more or less *general* ways of organising **counterfactual reasoning** about these possibilities, thereby enabling us to draw *specific* conclusions about how they would behave in any given set of circumstances. [Popper: Natural Selection]

A **functional schema** enables us to develop predictions by treating a system on analogy with **practical reasoning**. This works by allowing us to treat its **parts** as **means** in relation to the **whole** as an **end**. This lets us describe the causal role of the parts in terms of **success** and **failure**, and thereby to organise our reasoning about the causal relations between them in terms of the way failure **cascades** throughout the system. For example, the way that a narrowing of the aortic valve leads to heart failure, circulatory collapse, and ultimately death. The explanatory power of a functional schema thus lies in its introduction of the possibility of **malfunction**. Metzinger's theory provides us with three distinct types of functional schema: **unconscious drives**, **conscious systems**, and **self-conscious systems**.

Unconscious drives are causal systems that take *variable inputs* and produce *systematically*

correlated outputs. The **drive schema** enables us to understand the systematic correlation between these inputs and outputs, which I'll call **sensation** and **behaviour**, in terms of a single **goal state** at which the system aims. In essence, we treat sensation as information about the environment that is processed in determining how to behaviourally adjust it in line with the goal state. However, there need be nothing like a **representation** of the goal state playing a role within the system at this point, nor need the information involve an **internal distinction** between the system and its environment. These features emerge at the level of the **consciousness schema**, and the **self-consciousness schema**, respectively. The important point is that we can interpret a single system as containing a multiplicity of drives without thereby unifying these drives under a single end. In this case, we see the causal roles of the various drives as either **converging** or **conflicting** with one another in the production of overall behaviour, potentially **communicating** with one another, but never **prioritising** the various goal-states through some unified information processing activity.

It is the move to the consciousness schema that provides the representational elements necessary to unify disparate drives. This is because a **representational schema** allows us to extend a functional schema, through enabling us to develop predictions by treating a system on analogy with **theoretical reasoning**. This works by allowing us to treat its **internal states** as representing **external states** of its environment. This lets us describe the causal role of these states in terms of **accuracy** and **inaccuracy**, and thereby to organise our reasoning about the causal relations between them in terms of the way they contribute to the success or failure of the system's functional elements. The explanatory power of a functional schema lies in its addition of the possibility of **misrepresentation** as a reason for malfunction.

According to Metzinger, a conscious system is one that possesses what he calls a **Phenomenal World Model** (**PWM**). This is a functional subsystem that combines information from its various drives into a single store that is made **globally available** to them **simultaneously**. This sharing ties them together into a unified processing system with something like a common information format. This common format gives us the purchase we need to interpret the model as an **articulated representation** of the environment as a *whole*, with discrete *parts* that correspond to things *within* the environment. This is what underwrites Metzinger's claim that there is a **minimal neurological correlate** for any distinct representation present within phenomenal consciousness. It is also what lets us interpret the whole model as a representation of the **present**, insofar as different rates of information flow must all be integrated into a single window of processing. The final element of the consciousness schema is what Metzinger calls **transparency**, which means that the system must be

functionally unable to represent the possibility of **global misrepresentation**. Some of the elements of the model may become **opaque** to a degree, by incorporating representations of their own functional structure, and thereby the possibility of its malfunction, but this must never extend to the functional structure of the model as a whole. [explain the counterintuitive transparency/opaque metaphor: optical illusion example]

Metzinger essentially thinks that conscious systems have an ongoing **dynamic simulation** of the world, which they are structurally unable to recognise as a simulation. He sometime's refers to this as a state of **online dreaming**. Some might be tempted to oppose this to recent work in **embodied cognition**, which emphasises the extent to which our representation of the world is not self-sufficient, but requires embeddedness in the world to function properly. However, there is no real opposition here. All the PWM requires is a **minimal dynamic unity**, the elements of which can be threadbare in the absence of regular updates on the state of the environment. This is precisely what we see in **offline dreams**, where the level of detail and consistency of our representations is minimal, even if it is transparent to the point that we are often unaware of this at the time. For many things, we use our environment as a store of information about itself that can be dynamically accessed by interacting with it in various ways. The process of simulation can thus be **extended** even if its dynamic unity is **localised** in the brain.

The self-consciousness schema moves beyond unification by introducing a distinction within the PWM between the system itself and all other aspects of the environment. This constitutes a further subsystem which he calls a **Phenomenal Self Model** (**PSM**). The function of this subsystem is to move beyond the **unitary sensory processing** facilitated by the PWM to a **unitary behavioural processing**, which it achieves principally by enabling the system to represent not just the goal states it can achieve, but also its capabilities for achieving them. Metzinger is very explicit that PSMs can come in various degree of complexity, but he believes that there are three basic functional elements of human self-consciousness: **ownership** (the *incorporation* of PWM elements in the PSM) , **location** (the *perspective* of the PSM within the PWM), and **agency** (the *recognition* of events in the PWM as behaviour produced by the PSM) – each of which can fail in its own ways and thus be pulled apart from the other two [examples: body dysmorphia, phenomenal omnipotence] – and a more advanced functional element characteristic of higher states of human self-consciousness: **attentional agency** (the *control* of sensory input and processing).

Metzinger calls the system's self-representation the **Ego**. This is the **informational content** of the

PSM at a specific time, as opposed to the **functional role** it plays at all times. The human Ego contains various forms of information about us *qua* causal systems, ranging from more or less innate forms of **bodily information** about body position, integrity and possible action, to more or less acquired forms of **cultural information** about our social relationships, statuses, and possible negotiation. It also includes **mnemonic information**, converting the window of the present into a temporal tunnel, by enabling us to retain and recall our own history, and to simulate our possible futures. This is what Metzinger calls **The Ego Tunnel**. It is the integration of all these forms of complex information that turns our self-representation from a simple **body map** into a complex **personality**.

Metzinger's argument that there are no selves is fairly simple. The PSM's functional unity is not **persistent**: any aspect of it including the whole thing can be switched off an on at any time. This means that any persistence of the self between activations must reside in the content of PSM, or in the identity of the Ego. Metzinger then simply points out that there are no **criteria of identity** for individuating Egos corresponding to the intuitive self-understanding they involve. Insofar as the PSM is a part of the PWM, it is possible for it to misrepresent things in just the way the PWM does: it can fail to adequately represent the contours of the body, the historical episodes it has been involved in, or its capacities for action. However, the **phenomenal continuity** of ownership that it generates cannot misrepresent, because there strictly is no unity to represent independent of its own unifying function. The informational facets of personality that it tracks (bodily, cultural, mnemonic, etc.) can all vary independently of one another, making any criteria for binding them together more or less arbitrary. [example: radical neurosurgery]

3. Sellars: The Unification of the Subject

I'm now going to do my best to briefly sketch the way Metzinger's account can be extended by supplementing it with a Sellarsian **rational schema**. The real innovation of the rational schema is the introduction of a distinction between two types of informational unit: **sentential** (**S**) and **non-sentential** (**NS**). This produces a functional distinction between four types of information processes, based on how we pair inputs and outputs: **perception** (NS > S), **action** (S > NS), **inference** (S > S), and **coping** (NS > NS). The details of the rational schema are how we connect various subsystems capable of performing these different kinds of processing to form a system whose functioning is no longer merely *analogous* to reasoning, but which *actually is* reasoning. I'm not going to attempt to provide all these details here, but instead show how the issue of *responsibility* turns up within this

functional story.

The Sellarsian position is that a system is only reasoning once it is keeping track of relations between sentences, and being in this position involves these syntactic units playing a very particular functional role within its informational economy. We can program computers to process sentential inputs into sentential outputs, we can train parrots to produce sentential outputs (e.g., 'that is red') in response to non-sentential inputs (e.g., sensing a red object), and we can train dogs to produce non-sentential outputs (e.g., fetching a stick) in response to sentential inputs (e.g., 'fetch the stick boy!'). However, a rational agent capable of grasping the **meaning** of those sentences must be able to do all three of these things, and they must be able to do them in a way which connects them up properly. Sentences are syntactic units capable of playing all three roles within the same system (a state that can be observed, a state that can be inferred from or used to infer another state, and a state that can be brought about). The system needn't be able to process *every* sentence in all three ways (I can't observe 'several neutrinos have just passed through me', I (personally) can't derive the truth of 'the continuum hypothesis is independent of ZF set theory' from the ZF axioms, and I can't bring about the truth of 'Betelgeuse has gone supernova'), but it must be able to process *some* set of sentences in all three ways, and able to process *some* sentence-sentence transitions for *every* sentence. The **semantic content** of a sentence is the role that it plays within this **functional economy** of perception, action, and inference, but its **inferential role** is what ties this all together, by holding open the possibility that *any* sentence can be related to perception and action. [elaboration: taking-true Q on the basis of P, and making-true Q on the basis of P] It is the **capacity for inference** that transforms *sensation and behaviour* into *perception and action*.

The crucial thing about the capacity for inference is that it requires the ability to **dynamically track** one's *theoretical* and *practical* commitments, or to reliably keep score of the **claims** one is responsible for *justifying* and the **aims** one is responsible for *achieving*. This involves the ability to **dynamically update** one's commitments, by working out the **consequences** of existing ones, and revising them on the basis of **incompatibilities** between these consequences and newly acquired commitments. I'll call the functional subsystem that does this the **Rational Subject Model** (**RSM**). This subsystem in turn requires a capacity to dynamically track a broader space of possible **positions** than the ones one actually occupies. I'll call the functional subsystem responsible for this the **Inferential Space Model** (**ISM**). These two subsystems are functionally inseparable, insofar as they are two parts of the same dynamic process, but they exist in tension, insofar as the ISM must have the capacity to simulate other subject positions. [example: devil's advocate] The thing that

functionally distinguishes the RSM from these **simulated subjects** is its **reliability** in translating its commitments into actions. Something that could never act upon its commitments could never count as responsible for anything. It is in this sense that *rational subjectivity* is inseparable from *agency*.

I'm going to call the functional complex of the RSM and the ISM the **Core Reasoning System** (**CRS**). The crucial fact about the CRS is that it needn't be self-sufficient anymore than the PWM. It needs a minimal functional unity, but it can effectively store information in other internal subsystems, including Metzinger's PWM, and even in external environmental features, including other rational systems and the larger social systems they constitute. A rational system need not have a static store of sentences that it tracks and updates, as long as it can dynamically call up commitments and inferential relations between them on the fly. We can thus be *phenomenal consciousnesses* that **cope** with their environment by efficiently processing sensory inputs into into behavioural outputs without the need for reasoning, and nevertheless count as *rational consciousnesses* that **perceive** and **act**, as long as we have the capacity to channel this processing through the CRS when necessary. This is the functional essence of Kant's idea that *rational consciousness* is dependent upon the possibility of *rational self-consciousness*. I am only conscious of something insofar as it can be appended to the 'I think', which just means insofar as I can take responsibility for it as *mine,* or insofar as it can become an explicit commitment and be integrated with the others tracked by my RSM.

The question is now to what extent Metzinger's problem for phenomenal selfhood can be repeated for rational subjects. Although the account just provided places certain constraints on what can possibly count as a rational subject (e.g., this table cannot be a rational subject because it cannot track its commitments), these constraints still don't amount to **sufficient conditions** of individuation. There is no natural way of dividing up continuities in commitments and capacities into unique subjects any more than there is a natural way of dividing up continuities in personality into unique selves. However, there cannot be rational consciousness of any kind without the possibility of rational self-consciousness, and this means that it is a functional requirement that we have *some* way of individuating subjects, even if it does not demand any *particular* way. The crucial point is that whereas we can contrast the non-representational character of phenomenal continuity to the representational character of other aspects of the PSM, the same sort of contrast is not available between the commitments tracked by the RSM and the subjective continuity this presupposes, insofar as the former are **normative statuses** that are socially conferred by participation in the game of giving and asking for reasons. From this we can conclude that

subjective continuity is also a normative status conferred on rational systems by their involvement in this practice. This means that we must **socially institute** norms governing subjective individuation, even if these are to some extent arbitrary. [transcendental manifestation: the other example is the need for language] [personal identity: metaphysics vs. normative status]

However, our criteria are never completely arbitrary, because we are constrained by biological factors regarding the way in which we humans are functionally constructed to track our own states. This means that our individuation of ourselves as rational subjects is to some extent dependent upon our mechanisms for generating phenomenal selves. What this implies is that there is a **socially mediated functional relationship** between the CRS and the PSM. The domain of this social mediation is the **political realm**, wherein the massive information processing network of brains of which we are a part divides itself up into loci of *authority* and *responsibility*, before distributing specific authorities and responsibilities and between them. This makes the nature of **political agency** far more intricate than our phenomenal life might indicate.

4. Politics: The Menagerie of Agency

The crucial insight suggested above is that politics does not just involve the division of responsibilities between pre-existing subjects, but also the very individuation of the subjects that bear them. However, it also shows that this individuation does not operate without constraint. Nothing can be counted as a particular subject that does not have some capacity to track and act upon the specific commitments it is thereby responsible for. This is the basic form of the principle that **ought implies can**: one cannot have *any* responsibilities (or corresponding *rights*) unless one is capable of recognising and acting upon *some* responsibilities. There is much more to add to this point, but I will restrict myself to hinting at some of the issues that arise from the framework just proposed:-

i) Collective Agency: The mechanism independence of the rational schema implies that it can be applied recursively, so that it is entirely possible for rational agents to function as parts of other rational agents operating at a larger scale. This is what Hegel called **spirit**. This raises issues regarding how we *individuate* collective subjects (e.g., how we count systems such as corporations or states as responsible for their actions) and how we *construct* effective collective agents (e.g., how we constitute anything like **political will**). It also provides an example of the way that rational consciousness and phenomenal consciousness can pull

apart, at least insofar as whatever minimal dynamic functional unity a collective subject's representational systems could be said to have, anything that could be said to be its 'phenomenal inner life' would be nothing like our own.

ii) Overlapping Subjectivity: The possibility that the CRS can externalise aspects of its functioning implies that these externalised elements could potentially be shared by multiple distinct rational agents. The more one considers the **cognitive resources** required to engage in most reasoning tasks, the more likely it seems that **cognitive outsourcing**, and **cognitive division of labour** play an important part in the way we track our commitments and the inferential relations between them. This suggests that we may be able to discover and analyse the functional structure of a whole range of **socio-cognitive systems** that are not yet collective agents, and that these systems would play an important role in providing collective conditions of individual agency.

iii) Fragmented Subjectivity: This in turn raises the possibility that there could be systematic breakdowns of reliable tracking and action in relation to shifting external contexts (CRS networks) and internal contexts (RSM fractures). The reliable dispositions that are constitutive of agency could fragment, and in so doing fragment the subjects to which they are correlated. Just how much fragmentation can be permitted before this disrupts the identification of a given causal system with a given subject is a serious issue, and it forms a part of the larger question regarding criteria of subjective individuation. This fragmentation goes some way to providing more subtle explanations for *akrasia*, or **weakness of the will**, on both individual and collective scales.

iv) Systems of Subjectivation: The socially instituted norms for individuating subjects cannot simply prescribe the way existing causal systems are divided up into loci of responsibility, but must also prescribe the ways in which new causal systems are to be cultivated so as to be so dividable. This is the ordinarily thought as the political problem of **socialisation**, but it extends beyond the problem of the production and education of new generations of rational agents, to the systems there are in place for maintaining and stabilising the reliable dispositions of agents that are already socialised. These are what Foucault would call systems of **subjectivation**. This extends the study of socio-cognitive systems already proposed to include external factors that play roles in sustaining the minimal dynamic unity of the CRS itself.