COLLOQUIUM 2

What Kind of Theory is the Theory of the Tripartite Soul?

Rachel Barney
University of Toronto

Abstract

This paper discusses two related questions about Plato's account of the tripartite soul in the *Republic* and *Phaedrus*. One is whether we should accept the recently prominent 'analytical' reading of the theory, according to which the three parts of the soul are animal-like sub-agents, each with its own distinctive and autonomous package of cognitive and desiderative capacities. The other question is how far Plato's account so interpreted resembles the findings of contemporary neuroscience, given that this also depicts the mind as complex, partitioned, subject to conflict, and only very incompletely rational. The paper sketches the analytical reading, outlines the similarities and disanologies of the theory so understood to contemporary neuropsychology, and then steps back to consider three problems with such an interpretation. None is decisive; but they raise doubts as to whether the question of the title can really be answered in the way both the analytical reading and the modern parallel presume.

Keywords

Plato – soul – tripartition – psychology – neuroscience

Autobiographical Preface: The origins of this paper go back to the late 70's, when as a child I read Carl Sagan's pop science bestseller, *The Dragons of Eden*. Sagan was one of the first to make widely accessible the amazing new discoveries of neuroscience—especially Paul MacLean's theory of the 'triune brain,' which he likened to Plato's myth of the chariot in the *Phaedrus*.[1] I was fascinated by the idea that scientists were now rediscovering ancient Platonic wisdom about human nature (and by the mysteries it suggested—could that

1 Sagan 1977, 83.

© KONINKLIJKE BRILL NV, LEIDEN, 2016 | DOI 10.1163/22134417-00311P05

really be how science worked?): it may have been the source of my first interest in Plato. Over the years I waited to read a fuller, updated version of the comparison, but none appeared. When I organized a conference on the tripartite soul in 2005, I encouraged speakers to offer ancient-modern comparisons: there were no takers. My colleagues and I edited a book on the subject: again, no luck.[2] So in the end I had to do it myself. I discovered that it's far harder to draw a detailed, undistorted parallel than MacLean and Sagan made it look. Moreover, to do so requires addressing the enormous prior question of how to interpret Plato's theory in the first place. The comparison with contemporary neuroscience is made available by what I here call the *analytical* reading of Plato's theory, on which he is offering a natural-scientific analysis of the human psyche, as the locus of all our experience, cognition, and decision-making. So I begin in section I of the paper by setting out the analytical reading; I then proceed to the comparison, as best I can manage it, in section II. The rest of the paper then takes the form of a series of second thoughts: but is that *really* what Plato is trying to do? I consider three objections to the analytical reading which suggest that perhaps he is not offering the same kind of theory as contemporary science at all. These are, first, that the tripartite theory does not offer genuine causal explanations (section III); second, that it seems to invoke a 'meta' agent, a residual psychological power above and beyond the parts into which it analyses the soul (IV); and third, that it is oddly disconnected from Plato's theory of cognition (V). If these objections are true, there are fundamental ways in which the theory of the tripartite soul is not doing (and is presumably not meant by Plato to do) the kind of work that the analytical reading claims, or that any contemporary scientific theory aspires to. In the end my verdict is ambivalent: for me the title question remains an open and perplexing one. And that means that the resemblance of the tripartite soul to any contemporary theory is an open question as well.

I

In Book IV of the *Republic*, Plato famously divides the soul into three parts:[3] the appetitive part [ἐπιθυμητικόν], spirit [θυμός] or the spirited part [θυμοειδής],

2 Barney, Brennan, and Brittain eds. 2012.

3 Or so they are usually called in English, though Plato more often refers to them as εἴδη, 'forms' or 'kinds' (*Rep*. 435c1, c5, e2, 439e2, 440e9, 504a4, 572a6, 595b1, 612a5, cf. 580d3, 581c6; cf. γένη 435b5, 441c6, d3, cf. 441a1, 581c4; vs. μέρη 442b11, c5, 444b3, 577d4, 581a6, 583a1, 586e5). And he even more often avoids using any noun at all, as if to fend off the question of exactly what

and the rational part [λογιστικόν]. The rational part is home to our 'higher,' distinctively human capacities: objective calculation, abstract thought, and comparative evaluation. Its natural role in agency is as the provider of second thoughts and long-term planning; it corresponds to the Guardians in the state as the natural ruler of the soul, able to calculate what is best for the whole. As we might expect, the rational part develops slowly: children, and some defective adults, lack 'calculation,' its characteristic activity (441a). The spirited part is shared with other animals, and present in the young (441a–b). It specializes in the attitudes and emotions needed to regulate human (or more generally mammalian) social life. Spirit cares about honor, esteem, and status; its characteristic feelings include anger, pride, and disgust (439e–41c, 550b–3d). In the constitution of the psyche, its primary role is to apply those responses to reinforce and execute the decisions of the rational part; thus it resembles (and is dominant in) the Auxiliaries, the military class of the just city. Meanwhile the appetitive part is the primitive and dangerous locus of the basic and ineliminable physical drives we share with the other animals: for food, drink, and sex, as well as the desire for money. As such it is the counterpart in the soul of the 'productive' class in the just city—the farmers, traders and craftspeople whose job it is to keep the system physically functioning. Plato tends to present the appetitive part as irreducibly plural ('many-headed') and ineliminably primitive: even in the most educated and integrated psyche, it houses savage and lawless desires which find expression in obscene dreams (571a–2b).

Later Books of the *Republic* make it clear that it would be a mistake to view these parts simply as different functions or psychological powers. It is not the case, for instance, that reason = the power of thought, and the appetitive part = the power of desire. Rather, it turns out that *all* of them think, desire, strive, value, and experience pain and pleasure. This organic character of the parts is made vivid by Plato's metaphors, and in particular his image in Book IX of the parts of soul as different kinds of animal: the appetitive part is likened to a many-headed beast, the spirited to a lion, and the rational part to a sort of

---

these entities are. It is natural to wonder whether this indicates a weaker sort of individuation than 'part' would do, which would be a strike against the analytical reading. This is probably a misapprehension, however. As Burnyeat notes, for Plato an *eidos* is often a species understood as a *part* of a genus (in this case, presumably the genus 'soul')—so that *eidos*-language might be a version of part-language rather than an alternative to it (Burnyeat 2006, n21). It is perhaps preferred by Plato here because it suggests what *meros* does not, that the three parts are radically dissimilar in their causal powers and other features. (They are like the parts of a face, not 'parts' or pieces of gold, to apply a distinction drawn in the *Protagoras* (329d–30b).) In any case it seems to me unlikely that any reference to or echo of *the* Platonic Forms is intended.

miniature human being (588b–92b). As this suggests, the parts differ radically: but each is a kind of *agent*, with the complete, integrated package of cognitive and desiderative capacities that implies. (Thus the major operations familiar to folk psychology are distributed across them: no one part is the *sole* locus of perception, memory, emotion, belief, desire, pleasure, or *erôs*.) In a properly ordered soul, each will focus on using its distinctive capacities for the common good; but most souls are not properly ordered. To make clear the different ways in which souls can go right or wrong, Plato turns to his second family of preferred metaphors, taken from politics. The parts of soul are like classes—in the bad case, rival factions—within a city: they are constantly negotiating, manipulating each other, and struggling for power. Consider the soul of the oligarch, for instance:

> Don't you think that this person would establish his appetitive and money-making part on the throne, setting it up as a great king within himself, adorning it with golden tiaras and collars and girding it with Persian swords?
> … He makes the rational and spirited parts sit on the ground beneath appetite, one on either side, reducing them to slaves. He won't allow the first to reason about or examine anything except how a little money can be made into great wealth. And he won't allow the second to value or admire anything but wealth and wealthy people or to have any ambition other than the acquisition of wealth or whatever might contribute to getting it. (553c4–d7)[4]

The appetitive part is not very good at thinking, compared to the rational part: but it is smart enough to get reason to do the hard thinking for it, when it is in charge.

This understanding of the parts of soul as real, complex and to some extent independent sub-agents (rather than faculties, say, or mere psychological tendencies which Plato personifies for rhetorical effect) is the core of what I will call the *analytical* reading of the tripartite soul.[5] This line of interpretation is

4 Translations from Plato are from the many hands in Cooper ed. 1997, sometimes with revisions; in the case of the *Republic*, the translation is by Grube (rev. Reeve) 1992; for the *Phaedrus*, by Nehamas and Woodruff 1995.

5 The landmark argument for the analytical reading is Moline 1978. Major developments include: Cooper 1984; Kahn 1987; Reeve 1988. For particularly clear statements of the view, see Irwin 1995, 217–222; Bobonich 2002, 217–23; Lorenz 2006; Moss 2008; Brennan 2012. See also, albeit less explicitly, Annas 1981, 131–46.

defined by two complementary claims. First, the three 'parts' of the soul are, as I've just sketched them, robustly real and really distinct: each is a functional, agent-like unity defined by a distinctive package of cognitive, conative and affective capacities. Second, this division of the soul into parts is intended by Plato as a *full analysis* of human nature. The theory is one side of an equation, with the output of human thought, experience, and agency on the other side, and no remainder.[6]

The analytical reading represents one major family—perhaps the dominant one, at least in English-language scholarship—of interpretations of the tripartite soul in recent scholarship. I will refer to the non-analytical alternatives as *deflationary* readings. On a deflationary reading, Plato's representation of the 'parts' as agent-like is misleading: we are really only dealing with disparate tendencies or powers within a unified soul, whose autonomy he exaggerates for some heuristic, hortatory, or therapeutic purpose. Deflationary readings are hard to sum up, since scholars have a wide range of reasons for rejecting the analytical approach, and offer very different alternatives to it.[7] In any case my approach here will be to focus squarely on the analytical reading.

The analytical reading is solidly grounded in Plato's text; but only if we look well beyond *Republic* Book IV. It is not until Book IX that Plato emphasizes the way in which each part of the soul experiences its own desires and pleasures; it is here that the famous animal image is presented, as a kind of closing *tableau* to the whole account. And it is only in Book X that we get to see in any detail how the lower parts experience their own independent cognition. This comes in the discussion of the dangers of art—a passage whose importance

6 To be clear, the important claim here is that we are fully analyzable into subsystems *of this kind*, and *including* these ones; not that we consist in *only* these three. At *Republic* 443d7, in his peroration on the justice of the soul (quoted below), Socrates casually alludes to "any other parts there may happen to be in between." And at no point does he argue or assume that the three parts of soul distinguished in Book IV are the *only* ones there could be (even in the lazy heuristic way that he seems to assume the *kallipolis* must have exactly four virtues at 428a). Any claim to exhaustiveness is thus merely implicit and tentative. Presumably we are to ask ourselves, as the *Republic* proceeds, whether the psychological phenomena under discussion require postulating any further parts—and for Plato himself, the answer is evidently no. Likewise, though some scholars are concerned to establish that each of the three parts is not subject to any further subdivision (e.g., Lorenz 2006, Chs. 2 and 4; Irwin 1995, 218–22), I see no sign that Plato cares about this, or reason why he should.

7 Important deflationary readings include Shields 2001, 2010; Korsgaard 1999; Kamtekar 2006; Stalley 2007; Santas 2010, 82–88; Whiting 2012; and, in the end, Singpurwalla 2010. See also the objections to the analytical reading rehearsed at Annas 1999, 135–6; Bobonich 2002, 247–57; and Price 2009, 1–15.

for the tripartite theory is not always appreciated, though Socrates introduces it by announcing that the new discussion will benefit, as the earlier one in Books II–III could not, from the analysis of the soul (595a5–b1).[8] The passage will repeatedly be of importance for my discussion, so I will quote at length:

> Something looked at from close at hand doesn't seem to be the same size as it does when it is looked at from a distance.
> No, it doesn't.
> And something looks crooked when seen in water and straight when seen out of it, while something else looks both concave and convex because our eyes are deceived by its colours, and every other similar sort of confusion is clearly present in our soul. And it is because they exploit this weakness in our nature that *trompe l'oeil* painting, conjuring, and other forms of trickery have powers that are little short of magical.
> That's true.
> And don't measuring, counting, and weighing give us most welcome assistance in these cases, so that we aren't ruled by something's looking bigger, smaller, more numerous, or heavier, but by calculation, measurement, or weighing?
> Of course.
> And calculating, measuring, and weighing are the work of the rational part of the soul.
> They are.
> But when this part has measured and has indicated that some things are larger or smaller or the same size as others, *the opposite to it appears at the same time* [my emphasis].
> Yes.
> And didn't we say that it is impossible for the same thing to believe opposites about the same thing at the same time?
> We did, and we were right to say it,
> Then the part of the soul that forms a belief contrary to the measurements couldn't be the same as the part that believes in accord with them.
> No, it couldn't. (602c7–3a2)[9]

---

8 For a fuller discussion of this crucial text see Barney 2010.

9 The reading of this passage has long been controversial—but, I think, unnecessarily so. Nothing here is incompatible with the Book IV tripartition, or even much of an addition to it. As so often, Plato is simply proceeding on a 'need-to-know' basis, resulting in harmless ambiguity. *Whatever* refuses to accept the findings of reason cannot itself be rational, but must be something else—the appetitive part, or spirit, or both (or perhaps some

Socrates goes on to specify that this applies to the appearances generated by poetry as well as painting, as when somebody who knows better gives in to pain and grief.[10] So the lower, irrational parts of the soul have their own modes of cognition, ones which are necessarily crude and unreflective, but sufficiently like the judgements of reason for the two to conflict. Thus emotional turmoil and mental conflict are the expression of *cognitive* conflict: the parts of the soul feel and desire differently because they think about things—'see the world,' almost literally—in different ways. The spirited part will agree with Ajax that nothing is worse than the laughter of one's enemies; the appetitive part will agree with the erotic poet that nothing is more important than sex. Though we might not have guessed it from Book IV, Plato's theory thus gives a certain explanatory priority to cognition: our desires and intrapsychic conflicts will all ultimately be the expression of what and how our soul-parts think. [11]

The *Phaedrus* makes some important features of the *Republic* theory more explicit.[12] First, it frames it as the answer to a question about self-knowledge. Here is Socrates early in the dialogue, discussing the urgency of questions about the self:

> I am still unable, as the Delphic inscription orders, to know myself; and it really seems to me ridiculous to look into other things before I have understood that… . I look not into them but into my own self: Am I a

---

other part left unidentified, as mooted at 443d), as the case may be. See Moss 2008; Barney 1992, 286–7n5, and especially Lorenz 2006, 59–73, who notes that a number of back references in the vicinity show Plato to have both appetite and spirit in view.

10 Anyone who has read the *Protagoras* will be reminded of the discussion of *akrasia* at 352b–7e: it is the 'power of appearance' which leads us to wrongly choose a smaller good over a more distant larger one. One way to understand the account of the tripartite soul in the *Republic* is as Plato's answer to the question forced on us by the *Protagoras*: how can it be that our preference for the smaller good sometimes seems to *withstand* the information that it is smaller? Without the 'encapsulation' provided by the partition of the soul, and the potential for evaluative recalcitrance and conflict it provides, phenomena like mental conflict and *akrasia* are mysterious. The *Republic* can explain them in terms of conflicting evaluative beliefs, some of them resilient against certain kinds of new information or correction.

11 See, for example, Moss 2008.

12 For the purposes of this paper, I will take the *Republic* (at least up to the eschatological turn at 608c) and the *Phaedrus* to present the same psychological theory, which is the subject of my discussion here. The tripartite theory of the *Timaeus* is arguably very different (see, for example, Bobonich 2002 and Lorenz 2006), and the analytical reading captures it much less well.

> beast more complicated and savage than Typhon, or am I a tamer, simpler animal with a share in a divine and gentle nature? (229e5–30a6)

The passage points ahead unmistakeably to the famous myth later in the *Phaedrus* of the immortal soul as a chariot with reason as charioteer, drawn by a gentle horse (the spirited part of the soul) and a vicious one (the appetitive part). It also points ahead to a crucial passage in which Socrates discusses the correct method of philosophical or scientific analysis. To know any object is to be able to analyze it into its parts, and to know the natures of each part, that is, their powers [δυνάμεις] for acting and being acted upon.[13] Socrates considers the case of the soul in particular:

> Consider, then, what both Hippocrates and true argument say about nature [φύσις]. Isn't this the way to think systematically about the nature of anything? First, we must consider whether the object regarding which we intend to become experts and capable of transmitting our expertise is simple or complex. Then, if it is simple, we must investigate its power [δύναμις]: What things does it have what natural power of acting upon? By what things does it have what natural disposition to be acted upon? If, on the other hand, it takes many forms [εἴδη], we must enumerate them all and, as we did in the simple case, investigate how each is naturally able to act upon what and how it has a natural disposition to be acted upon by what. (270c9–d7)

The methodological importance of this passage is hard to overstate. Together with the discussion of collection and division at 265c–6d, it tells us a great deal about Plato's mature conception of philosophical (or scientific) method. To understand something is to be able to analyze it into simple and complex components, and identify their causal powers. The case in point is the soul, and it is telling that Socrates speaks here, as in the *Republic*, of the components of the soul as 'forms.' Here in the *Phaedrus*, Socrates' immediate concern is with rhetoric, and the knowledge of the soul which it requires; and the rhetorician must know how many forms of soul there are so as to know how they are configured in people of different types: "hence some people have such-and-such a character and others have such-and-such … People of such-and-such a character are

13 Cf. *Soph*. 247e for the power to act and be acted upon as the mark of being. Note that there, powers are possessed by forms: this suggests that deflationary readings on which the parts or 'forms' of the soul are *themselves* merely powers are making a category mistake.

easy to persuade by speeches of such-and-such a sort …" and so forth (271d3–6).

Plato here seems to be giving us some retroactive directions on how to read the theory of the *Republic*, and they are friendly to the analytical reading. First, his account there of the 'forms' of the soul is to be taken as a scientific analysis into simpler components defined by their causal powers, analogous to medicine's analysis of the body. Second, the configuration of these parts and powers in an individual is what determines his 'personality type,' as manifested in his being receptive to this or that sort of rhetoric; to be able to apply the general analysis to the individual and spot distinctive features and peculiarities is also a matter of scientific psychology (again, compare the doctor). And third, such a scientific analysis of my own individual psyche would give me *self*-knowledge: for to know oneself is to know the nature of one's soul in all its complexity. I will from here on take the analytical reading as incorporating all these claims. The 'self' is of course a slippery concept, and we need to be careful about retrojecting modern assumptions about it. But what Socrates has in mind in the *Phaedrus* seems familiar, and easy enough to specify in a loose and pretheoretical way. My self is what I refer to by 'I' and 'me,' the subject of my experience and the agent responsible for my behavior—who I am as a thinking, acting, experiencing person. (Of course, whether there *is* a single, unified self, picked out in the same way by all these roles, is itself an enduring philosophical question.) On the analytical reading, then, the soul just is its three parts; and the self just is the soul.

In the rough and general form I've now set out, the analytical reading leaves many questions untouched. For instance, how exactly do the parts communicate or fight with each other? How does one part manage to seize power or to impose its values on the rest? To what extent can each be transformed by its circumstances and cultivation? Different versions of the analytical reading give different answers to questions like these; so perhaps it would be better to speak of a family of readings, or a general interpretive strategy. But every version will differ more deeply still from deflationary readings which take Plato's presentation of the parts as agent-like to be merely metaphorical, heuristic, or a rhetorical exaggeration.

II

So what *kind* of theory is this? A very familiar and modern-looking one, or so it seems. On the analytical reading, as Socrates' invocation of Hippocrates in the *Phaedrus* brings out, Plato's aspiration is to an empirically adequate natural-

scientific analysis of human nature, one which can provide a framework for explaining our behavior, individual differences and lived experience.

To see what I mean by 'familiar,' consider the sort of analysis of the human mind offered by contemporary neuropsychology.[14] Here too, we turn out to be not one but many, and recent progress in understanding human nature has come from understanding the subsystems into which we are divided.[15] Those subsystems range from the most 'primitive' (automatic, impervious to reflection, shared widely with other animals) to the fully rational and distinctively human. The rational, conscious self with which we like to identify turns out to be a small part of the whole—the tip of 'the unconscious iceberg'—and only very imperfectly in control.[16]

Some theorists of the mind have worked out the analysis of these subsystems in strongly Platonic terms. On Paul MacLean's influential theory of 'the triune brain,' our cognitive architecture consists of three main structures: the rational, distinctively human *neocortex*, home of self-control and abstract thought; the *limbic system*, shared by other mammals and responsible for emotions; and the *R-complex* or *reptilian brain*, shared far more widely and responsible for more basic physical responses. The theory is obviously and self-consciously Platonic (though there are still some mismatches).[17] However, the theory of the triune brain has fallen out of favor in recent decades.[18] In particular, the idea of the 'limbic system' seems to have dissolved. It is now widely seen as at most a conveniently nebulous *façon de parler*, rather than capturing a real unity.[19] Meanwhile the cognitive structures involved in emotion, initially

14 An alternative analogue, which I cannot here explore, is Freud's tripartition of the psyche into ego, superego, and id. For comparisons, see Ferrari 2007; Santas 1989; and Price 1990. Whether the Freudian model is particularly friendly or otherwise to the analytical reading is unclear to me.

15 For the following survey, see for instance Gazzaniga, Ivry, and Mangun 2014; and Carter 1998. For the now-commonplace finding that we are not one but many, see especially Dennett 1991; Gazzaniga 1985 and 2011; Ramachandran 2011; and Ornstein 2003: "There is no single mind but many; we are a coalition, not a single person… . We are unaware of how we decide and even 'who' is deciding for us" (21).

16 Gazzaniga 2014, 78; cf. pp. 66, 102.

17 See MacLean 1973 (also 1990); for a riveting if now-outdated popularization, see Sagan 1977.

18 See Goldberg 2001: "the so-called limbic system, a somewhat outdated construct implying a functional unity among these structures, whose heuristic value has been increasingly challenged" (31). For a state of the art 'textbook' verdict see Gazzaniga et al. 2014, 428–9.

19 For instance, "it has been impossible to establish criteria for defining which structures and pathways should be included in the limbic system" (Gazzaniga et al. 2014, 429).

for the most part attributed to it, have turned to be far more complex than anticipated.[20]

If the theory of the triune brain were correct, the psychological theory of the *Republic*, analytically read, would not only be boldly original, insightful, and suggestive; it would have some claim to the more vulgar virtue, to which Plato so rarely stoops, of being empirically true. As it is, more recent scientific work suggests that any resemblance between ancient and modern holds only at a rather high level of abstraction—and of course there are some obvious objections to drawing any analogies at all. To begin with the most obvious, these are theories of, on the face of it, different things: on the one hand the brain, understood as a part of the body and subject to the same general physical laws, and on the other hand the soul, which Plato regularly *contrasts* with the body. This difference does not cut so deeply as it might appear, though, since plenty of contemporary scientists and philosophers are happy enough to take the brain as equivalent to the mind, and the mind as the locus of soul and self. (Virtually all the pop science works cited in this section move seamlessly from talk of brain structures to talk of human nature, the explanation of behavior, and even first-personal experience.) Since the *explananda* turn out to be much the same in the two cases (and since nothing in the *Republic* theory *excludes* a fully 'physical' realization of the soul, which Plato himself arguably supplies in the *Timaeus*), the difference between 'physical brain' and 'non-physical soul' seems not to be very salient.[21]

---

20 See Damasio 1994 and LeDoux 1996. The question of how to relate modern thinking about the emotions to the Platonic soul is a particularly difficult one: it is far from obvious that 'emotion' is a category operative in the psychology of the *Republic* at all.

21 However, the difference between a physical and a non-physical account does mean we need to be on guard in our comparisons: we cannot assume that a single 'location' in the Platonic soul must correspond to a single 'location' in the brain. For instance, Plato locates anger in a single part of the soul, the spirited part. If we ask whether contemporary neuroscience reaches a similar result, the question should not be whether anger is always located in a single brain structure. Rather, the question must be (1) whether Plato's 'anger' turns out to be a single thing, or dissolves on scientific scrutiny; (2) whether that one thing is realized, as he claims, by *quasi-rational* structures (of the kind that are shared by other mammals, neither reptilian nor distinctively human); and (3) whether those structures are *also* responsible, as he claims, for shame, the drive for honor, etc. More generally, for there to be a neuropsychological counterpart to Plato's spirited part is simply for *some* unified grouping of structures or pattern of circuits, however complex or widely distributed, to reliably instantiate the cluster of processes and features he attributes to spirit. To be clear, I do not think that there *is* a neuropsychological counterpart to anything as complex yet unified as a Platonic soul-part—not least because of the failure of the 'limbic system' to capture a real unity.

A more problematic difference is that Plato's subsystems seem to be of the wrong kind—on the wrong scale—to be genuinely explanatory by modern lights. That is, the cognitive 'parts' salient for scientific explanation these days are not at the same level of complexity as Platonic soul parts. For instance, we can now identify at least thirty-two distinct systems involved in visual processing alone, organized into complex feedback loops. (There even seems to be a distinctive area in the brain devoted to the recognition of fruit.)[22] The attention of scientists has, it seems fair to say, largely been devoted to identifying such micro-structures in the brain, and showing how even simple-seeming processes must be broken down into surprisingly complex components—*not* to identifying the kind of large-scale, agent-like clusters of multiple processes which Plato identifies as 'parts' of soul. The *Phaedrus* of course *exhorts* the scientific psychologist to proceed in analysis all the way to the simplest components and their causal powers; but the modern comparison shows how very far Plato himself is from doing so. Whether there is a really deep difference here in the conception of the soul—whether contemporary neuroscience *excludes* the sort of large-scale patterns of agent-like psychic organization which Plato identifies—or just a difference in focus and level of explanation, is impossible for me to say. But it seems safe to say that no unitary system as complex and agent-like as the spirited part, say, has so far been discovered and vindicated as genuinely explanatory.[23]

---

22 Gazzaniga 2011, 40.

23 Given all these reasons for caution, we might be tempted to look elsewhere for analogues in modern psychology (see also footnote 14 on Freud above). In 'Dual Systems in 400 BC: Plato's Parts of the Soul and Contemporary Psychology' (ms), Jessica Moss discusses an alternative contemporary analogue for Plato's theory, namely, the accounts of human reasoning in terms of 'System 1' and 'System 2' put forward by Kahneman and Tversky and others in recent decades. (The resemblance is also noted by Singpurwalla 2010, 890.) (See Kahneman 2011 and, for a recent and balanced survey of the state of the art, Evans and Stanovich 2013). Moss is right, I think, that this comparison can help to illuminate Plato's conception of rationality: the sort of irrationality exhibited by the lower parts of the soul consists not (or at any rate not *essentially*) in a lack of propositional or conceptual capacity, but in its being restricted to unreflective judgements of the System 1 type. The limitation of the analogy is that it is far from clear, and indeed highly controversial, what these 'Systems' or processes themselves *are*. Stanovich, who introduced the terms 'System 1' and 'System 2,' now prefers to speak in terms of 'dual processes' for just this reason, and seems skeptical that there is any real unity to System 1 (Evans and Stanovich 2013, 225). Recent critiques of dual process theories (helpfully summarized in Evans and Stanovich 2013) raise enormous difficulties for any understanding of the 'processes' or 'systems' which would reify them, treating them as unified, stable, causally efficacious psychic entities. (It seems to me, for what it is worth, that the contrast between Systems 1 and 2 is best

Still. Even with all these reservations, disanalogies, and grounds for caution duly noted, Plato—as understood by the analytical reading—seems to have been on to something deeply important. He seems to have been right to treat intrapsychic complexity and the interaction of semi-autonomous, radically different psychic parts as the explanatory key to human behavior; right to suppose that those parts can be located on a spectrum from rational to not at all so; and right that the rational control of the irrational is typically very imperfect indeed. At a minimum, he presents an ur-version of what now looks to be the right *kind* of theory. We can outline the essential features of that shared theoretical project as follows:

1. Knowledge of human nature is to be gained through the empirical natural-scientific study of the mind or soul (in a word, psychology), which forms either a branch of medicine or a close counterpart to it.
2. That study, like the study of any complex whole, takes the form of an *analysis*. The job of scientific psychology is to identify the components of our psyche at various levels of complexity and their causal powers; to explain their interactions; and in doing so account for our behavior, lived experience, and individual differences.
3. This analysis discloses a multiplicity of *subsystems* or modules.[24] These modules are to some extent *insulated* in their functioning from each other, and some are irremediably primitive and non-rational. Many have homologues in other mammals and some are more primitive still ('reptilian'). The most primitive of these modules house desires linked to the needs of the body. Several levels of non-rational cognition are absolutely necessary for normal human functioning.
4. In a healthy psyche, the non-rational modules are regulated by a rational *command center* which is fully developed only in adult humans; this is at once the locus of self-control, practical reasoning, and abstract thought and computation. It is characterized by distinctive motivations as well as high-order reasoning. The construction of an ongoing unitary self obedient to conscious goals and beliefs is a major function of this center.

---

understood as an analytical tool deploying essentially relative concepts, like Aristotelian matter and form.) In short, they do not seem to be like Platonic soul-forms or -parts at all, at least as the analytical reading presents them.

24 The *locus classicus* for the concepts 'modularity' and 'encapsulation' is Fodor 1993; for a range of more recent permutations and reservations, see Faucher and Tappolet 2006.

5. Our cognitive parts have distinct functional *specializations*, but overlap enough to *conflict*. We can see this from the way a primitive or narrowly specialized module may persist in some stance in opposition to the findings of reason ('encapsulation'): optical illusions, for instance, or self-destructive or addictive impulses. It is not unusual for the rational command center to lose out in a conflict with an irrational module; and often its work includes confabulation or rationalization in service to a non-rational demand.
6. This analysis discloses that the unitary *self* is an illusion (or at best a construct, and a small part of the story, as per (4) above). To answer Socrates' question in the *Phaedrus*, we are each of us not one but many. This is true not only in the weak sense that we are complex wholes, but in that the subsystems within us have a certain autonomy and their integration is imperfect. There is no unitary 'I,' over and above the parts.
7. The central categories of folk psychology—*complex processes* like perception, memory, emotion, desire, pleasure, decision, action—cut across this analysis in complicated ways. All must themselves be distinguished into subspecies, and each of these into component processes; some (very likely including 'emotion') may not be natural kinds. None has a simple, unitary 'location' in just one module.

This seems to me to add up to rather a lot. Just what we should make of it—for instance, whether and how it should affect our *evaluation* of Plato's theory (or for that matter the modern one)—is an interesting and tricky question, but not one I will here pursue.[25] For my purposes, the point is just that we now know (or think we know) in a general way what a plausible natural-scientific analysis of the human psyche looks like. And so if we accept the analytic reading of Plato, we have a ready answer to the question of my title: he was doing roughly what neuropsychology is trying to do today, and doing it with astonishing prescience. What this does not do, of course, is tell us whether we *should* accept the analytical reading; and this is the question to which I now turn.

In the following sections I will discuss a few serious problems for the analytical reading. I must warn that these are the ones which worry me the most, rather than being a survey of all the difficulties that have been raised in the literature. In particular, I will have little to say (except right now, by way of

25 Katja Vogt's comments on this point seem to me to ask all the right questions; I am not certain of what my own answers would be. As the following paragraphs indicate, I do think that such parallels may at least be helpful in suggesting how the Platonic position can be defended in response to objections.

dismissal) about the concern that has probably done the most to motivate deflationary readings. This is a worry about psychic *unity*—about how the soul under tripartition still remains one. For the modern parallel can help us to see, I think, that the analytical reading should happily dismiss this as a pseudo-problem. In truth deflationists seem to be motivated by two worries about unity, of very different kinds. One set has to do with Platonic texts which suggest that the multiplicity of the human soul is only superficial. The argument for immortality and the 'sea-god Glaucus' passage of *Republic* X suggest that in some sense the 'true,' immortal soul is unitary (608c–12a); this is also what we would expect from the *Phaedo*, and it seems to be the view of the *Timaeus* as well. However, in *Republic* X Plato himself flags the point that there is a big difference between studying the incarnate and the discarnate soul: the 'sea-god Glaucus' passage amounts to the suggestion that we turn to study the soul in its pure, discarnate state (611b–2a6). Now there is, I take it, a real philosophical problem about how the incarnate and discarnate human souls are to be related; and Plato's solution to it is rather unclear. I am inclined to say that he *has* no firm solution, and that *Republic* X, the *Phaedrus* myth, and the *Timaeus* account of incarnation—which are themselves notoriously inconsistent—represent a succession of more or less unsatisfactory attempts to wrestle with the problem. Be that as it may, we are not licensed to use what Plato says about the soul in eschatological contexts as a constraint on his empirical psychology of incarnate humans. He himself warns us that these are two separate topics, and that the *Republic* has treated only the latter.

The other worry about unity is purely philosophical: it is, to put it roughly, that explaining how the soul forms a unity is a constraint on any adequate psychological theory. If Plato's theory does not explain how we are one *as well as* many, his theory fails. Here I think the modern parallel can help us to answer on behalf of Plato (analytically read): he should flatly reject any such constraint, with some of the bullet-biting glee of his modern neuropsychologist counterparts.[26] We are as unified as the best empirically defensible account shows us to be, no more and no less. And that turns out to be not very unified at all: "the self is not the monolithic entity it believes itself to be" (Ramachandran 2011, 247). *Tant pis*.[27]

---

26 Outstanding examples include Gazzaniga 2011, and Dennett 1991. Such theories can still allow that a certain kind of unity may be a psychological *norm* to aspire to, as Plato also holds (*Rep*. 443e2).

27 A more precise and sophisticated version of this worry in the end meets a similar fate. This is that, as Price 2009 has very clearly brought out, Plato's theory seems unable to make sense of our experience of a unitary *consciousness*. Surely on the analytical reading

The three problems for the analytical reading which follow are ones which I find less easy to dismiss and more fruitful to worry about. By the same token, I am not very happy with my answers to them. To anticipate, I do not think anything ahead counts as a *decisive* objection to the analytical reading. But we will see that Plato's account so understood remains puzzling; and this does suggest that the analytical reading may fail to capture some of his basic theoretical aims.

III

One familiar objection to the tripartite theory, analytically read, is the *homunculus* problem. As Julia Annas first put it, invoking Daniel Dennett, psychological explanations in terms of sub-agent-like mental entities—'a little man' inside the big one—often just look like a pseudo-explanatory regress.[28] But, Annas notes, this is not really a worry so long as the sub-agents are *simpler* than the whole, rather than fully replicating and redundant with it. One might fear that the rational part in particular is so anthropomorphic as to be redundant and unexplanatory; but even this is not quite right, as Annas points out, since the values of the rational soul and the self as a whole may come apart (Annas 1981, 145). Actually one can put the point more strongly: the rational part would only *really* be like a miniature human being if it contained within it miniaturized versions of *all three* parts, in a horrible regress reminiscent of Anaxagorean physics. The rational part is like a human being in that it is the locus of what is *distinctively* human—not in that it reproduces everything that we are.

Annas's dismissal of the homunculus objection is, I think, right in a general way; and interpreters should beware of throwing around the term 'homunculus' as if the very word conjured up a powerful *reductio* of the analytical reading. However, there is a variant on the objection worth exploring. This is that the account is still unexplanatory in that it reproduces and leaves unanalyzed the

---

we would expect the Platonic soul to house three more or less independent centers of consciousness; and that is hardly how we experience ourselves. Here again I think the analytical interpreter must simply deny the alleged explanandum. In fact, and baffling as it may be, the *Republic* seems to have nothing to say about consciousness *at all* (and note what looks like a rather laborious discovery of the concept at *Philebus* 33c–4a). And from the standpoint provided by the counterpart modern theories, this is perhaps just as well. For it turns out that consciousness is just 'the tip of the psychic iceberg,' which it largely misrepresents—not the authentic interior monologue of a unified self. (See Gazzaniga 2011, esp. 66, 78, 102; and Ramachandran 2011.)

28 See Annas 1981, 143–6; Bobonich 2002, 221–3.

operations we typically ascribe to the agent as a whole. For instance, we might think that a useful, genuinely scientific explanation looks something like this:

> The human being feels fear when circuits in the amygdala fire in response to a stimulus from the thalamus or hippocampus.

Now compare:

> The human being feels fear when the amygdala feels fear.

Here something does look defective about the second 'explanation.' It is not that the first explanation gives us a *definition* of fear; but it does, as is typical of physicalist explanations, tell us something about the 'how,' the physical realization of the operation, which seems genuinely informative. The second sentence gives only a source or *location* for an operation which is itself left unanalyzed. So a more telling version of the homunculus objection would be that the explanations provided by the tripartite theory are merely 'locationist':[29] the human feels hunger when the appetitive part feels hunger, cares about honor because the spirited part cares about honor, believes the Pythagorean theorem because the rational part believes it, and so forth. The question, then, is whether locationist explanations can be genuinely explanatory—given that, so to speak, they leave the verbs untouched.

In response I would suggest, tentatively, that the answer is yes; and that this objection helps to clarify what Plato is up to with the tripartite theory, rather than showing it to fail. By way of analogy consider:

> Ruritania invaded Carpathia because the King was angered at its Grand Duke.

or:

> Neutralia banned all newspapers because the Minister of the Interior wanted to silence the *Daily Journal*.

or:

29 This term is used in neuropsychology slightly differently, for explanations which attribute some process to a *single* location in the brain (e.g., fear to the amygdala, or all emotions to the limbic system): see Gazzaniga et al. 2014, 429ff.

> Oceania is unreliable about international trade agreements because of conflicts within the Central Committee.

In explanations like these the behavior of a complex whole is explained by reference to the desires, power relations, and agency of one or more of its parts. Explanations of this kind may indeed usefully answer a why-question. They are especially appropriate for explaining the behavior of complex 'political' systems whose inner workings are not directly observable—which is of course just what the Platonic psyche is. For that matter, it may even tell us something useful to say that fear takes place in the amygdala: it directs us, for instance, to explain anomalous fear responses by checking whether the amygdala of the agent is damaged. Such explanations do not tell us what it *means* to invade a country or sign a trade agreement—it is assumed that we know that, or that the question can be set aside for now—but only how such a decision came to be made by the agent as a whole. So we just need to be clear that the tripartite analysis of the *Republic* is not intended to reveal what (for instance) a belief, desire, pleasure, or a decision *is*: such questions are held in reserve for later dialogues. Setting them aside, locationist explanations can still be genuinely helpful—for instance, when we explain the paradoxical and puzzling behavior of a Leontius by reference to the conflicting desires of his soul-parts (*Rep*. 439e–40a).

However, this does seem to be a disanalogy with the comparable contemporary theories, which are more ambitious. It is not just that current research favors explanations which involve distributed functioning over multiple brain structures. The more important point is that these more complex explanations go with subdividing the process in question into the distinct contributions made by each structure—so that we do begin to get an analysis of *what happens* in fear, vision, and so forth. Very occasionally Plato's tripartition provides for something like this. For instance, we could say that the *inhibition* of a shameful erotic impulse occurs when a certain kind of unnatural desire 'pulls' the appetitive part; the rational part registers this, determines itself to oppose it, and relays instructions to the spirited part; spirit, experiencing shame, sends an affectively vivid message (involving fear of painful punishment, perhaps) to the appetitive part; and the appetitive part backs off. But this still leaves the more elementary processes of desire, shame, pain, fear etc. unanalyzed. At most, the theory of the tripartite soul gestures towards a research program in the analysis of psychological processes and functions.

IV

Here is another puzzle for the analytical reading. At the end of his account of soul and virtues in *Republic* IV, Socrates sums up the psychic state of the just person as follows:

> One who is just does not allow any part of himself to do the work of another part or allow the various classes within him to meddle with each other. He regulates well what is really his own and rules himself. He puts himself in order, is his own friend, and harmonizes the three parts of himself like three limiting notes in a musical scale—high, low and middle. He binds together those parts and any others there may be in between, and from having been many things he becomes entirely one, moderate and harmonious. (443d1–e2)

Now just who, we might wonder, is doing what to whom here? Who is this unexplained *he* who organizes the parts of *his* soul in this way? After all, the argument of the *Republic* so far would naturally have led us to believe that—as the analytical reading makes explicit—the human being as a psychological agent simply *is* the three parts of the soul which Socrates has just now been at pains to distinguish.

This mysterious passage suggests that there is indeed a self—a *me*, a psychic agent—which is something over and above the parts of my soul and is able to act on them.[30] This vaporous self turns up again in Book IX, when, in a passage probably meant to recall this one, Socrates urges that we should do everything to preserve the rule of the rational part (588e–91e). And this shadowy superordinate agent is not only invoked in contexts of exhortation. The oligarch too is described as behaving like a distinct agent in relation to the parts of 'his' soul, in the passage I cited above in section I. 'He' sets his money-making part on the throne; it is 'he' who won't allow reason and appetite to value or deliberate about anything other than the acquisition of wealth (553c–d). The agent here seems to be identified with the oligarch himself; yet he is distinguished

30 Irwin 1995, 255–8 notes the problem in relation to the oligarch, and identifies the three interpretive options (a *façon de parler*; an invocation of a further agent; or an allusion to the agency of some previously identified part of the soul). His preference is for the last of these, seeing an allusion to the rational part; I reach a similar conclusion at the end of this section, but without much conviction and for rather different reasons. For a contrasting view, see the 'power struggle' reading of *Republic* VIII–IX presented by Johnstone 2011.

from the parts of the soul, on which he acts. So the analytical reading must be wrong.

Call this the problem of the *extra* or *meta-agent*.[31] We might try to dismiss Plato's invocation of the meta-agent as just a sloppy (but forgivably convenient) *façon de parler*—a shorthand for the power relations among the parts. Plato's exhortations then amount to: 'Parts, order yourselves as follows!' And his depiction of the oligarch is just an account of how the parts sort themselves out in his case: "a picturesque but eliminable feature of Plato's exposition, not to be taken literally as part of the explanatory model."[32]

But now consider the crucial 'function' argument of *Republic* Book I. Here Socrates first proves (in at least a preliminary way) that justice, as the virtue of the soul, is necessary for happiness. The mainspring of the argument is an inductive argument, over horses, sense organs, and tools, for the conclusion that the function of a thing is "that which one can do only with it or best with it" (352e4). Whatever has a function has a virtue or vice, by possessing which it performs its function well or badly.

> Come then, and let's consider this: Is there some function of a soul that you couldn't perform with anything else, for example, taking care of things, ruling, deliberating, and the like? Is there anything other than a soul to which you could rightly assign these, and say that they are its peculiar function?
> No, none of them. (353d3–8)

The soul is here treated as a tool which 'you' use—and use for deliberating and ruling, so that the rational part must be included in the soul which is here distinguished from its user. Admittedly the 'instrumental' dative locution need not be used for a tool wholly other than its user: it may signal merely the part

---

31 I take the helpful term 'meta-agent' from Christoph Horn's paper, 'Plato's Republic IV: Autonomous Parts of the Soul?', delivered at the 2015 conference, 'Plato's Other Souls,' at Ruhr-Universität Bochum, organized by James Wilberding and Jana Bleckmann. All the papers at this conference were highly relevant contributions to the debate over the analytical reading and its alternatives, but none resolved what seemed to me the central *aporiai*, and I wanted to avoid frequent reference to as yet unpublished work; so apart from stylistic revisions I have chosen to leave this paper in the form delivered at BACAP in 2014. An earlier version of the paper was delivered at the Northwestern University Conference on Plato's Psychology, March 2008 (my thanks to Richard Kraut and the other conference participants for discussion there) and a much-revised one at USP in 2014 (likewise Marco Zingano and his ancient philosophy seminar).

32 Kahn 1987, 82n8. Cf. Bobonich 2002, 531n27: "occasional loose language".

of the agent which performs some operation.[33] Still, there is at a minimum a distinction being drawn between self and soul. Worse, a passage of Book X suggests that this instrumental relation may have philosophical significance for Plato. Here he affirms that virtue is conceptually dependent on use:

> Then aren't the virtue or excellence, the beauty and correctness of each manufactured item, living creature, and action related to nothing but the use for which each is made or naturally adapted? (601d4–6)

So if the general principle here applies to the soul, its capacity for virtue and vice depends on its having a use; and if 'use' is to be understood on the model of the use of tools by craftspeople, the Platonic self *must* be something over and above the soul, and distinct from it as user is to used.

This distinction between self and soul is not restricted to dialogues in which tripartition appears. In the *Apology*, Socrates describes his ethical mission as follows:

> I go around doing nothing but persuading both young and old among you not to care for your body or your wealth in preference to or as strongly as for the best possible state of your soul, saying to you: 'Wealth does not bring about excellence, but excellence makes wealth and everything else good for men, both individually and collectively' (30a7–b4).

Here too the self being exhorted seems to be an agent over and above the soul, to which it is urged to attend.

Now all this is rather disconcerting. For this meta-agent is never properly labeled, explained, or related to the parts of soul, some of whose functions it seems to duplicate. Moreover, we have reason to think that it is un-Platonic. In introducing the analytical reading, I noted that in the *Phaedrus* the analysis of the soul seems intended to answer Socrates' quest for self-knowledge. Moreover, the *First Alcibiades* contains a powerful argument to the effect that the self *is* the soul (128a-130e). The *Alcibiades* articulates quite a rich conception of the self, as at once discursive, social, and practical: the 'I' is what speaks and is spoken to by other selves, and what acts using my body. And only the soul—here left an unanalyzed place-holder—could be suited for that role of self. Of course, the authenticity of the *Alcibiades* is deeply contested, in a controversy

33 This locution in turn is taken to license putting the part or tool in the subject position as if it were the agent: hence we are to be wary of using it for the sense organs in the case of perception, for the purposes of Plato's argument at *Tht*. 184–7 (cf. Burnyeat 1976).

which will likely never be resolved.[34] Indeed it's easy to imagine it being written by a follower of Plato as an *hommage* and a clarification of Plato's views on precisely this point—even as an implicit *correction* of the 'meta-agent' passages I've cited. But if so, I think the anonymous author gets it right: this *should* be Plato's view. Just think, for instance, of the whole argument of the *Phaedo*, or the Myth of Er at the end of the *Republic*, or any other Platonic account of immortality. All assume that my soul will retain the traces of what I have learned and done, and can be justly punished for my crimes; and above all, that its survival is sufficient for my own. None of the proofs of the immortality of soul in the *Phaedo* would make my death any less frightening unless my soul is *me*.[35] Moreover, Plato is consistent in insisting (not least in some of the very passages which seem to invoke an extra agent) that the health of my soul is the greatest good for me, and decisive for my happiness; exactly why this is so is left surprisingly unclear, but the easiest and most obvious explanation would be that my soul is *me*.[36]

So how should we understand the passages which seem to invoke a meta-agent? A clue is perhaps to be found in the very fact that it is always an *agent*, and active in relation to the soul itself. So the crucial question turns out to be: granted that these passages feature a distinction between psychic *agent* and *patient*, how much of a real distinction does that imply? In other words: if a *reflexive* action is one in which a single entity is both agent and patient, does Plato deny that real reflexivity is possible? For only in that case would the meta-agent *have* to be something more than a convenient *façon de parler*.

Now there is some evidence that Plato does find reflexivity paradoxical, and perhaps even impossible. In *Republic* IV, Socrates' discussion of the virtue of *sôphrosunê* begins with the complaint that talk of 'self-control' is, on the face of it, ridiculous. The problem is then dissolved by the tripartite analysis: really, a soul or city is self-controlled when the better part in it controls the worse.[37] Plato's enthusiasm for this solution does not show that reflexivity is a downright impossibility (as opposed to superficially puzzling). On the other hand, it is striking that in the *Charmides* and other early dialogues, this obvious understanding of *sôphrosunê* as self-control is never even mooted; this suggests that

34 Notwithstanding the best efforts of Denyer 2001.

35 Or, as Evan Keeling has pointed out to me, at least *the most important part* of me. I am not sure how to make this weaker view precise, though, and see no good grounds for preferring to attribute it to Plato.

36 Cf. *Apology* 29d–30a, *Gorgias* 477b–e, *Crito* 47d–8a.

37 It would seem to dissolve the paradox equally well if the worse part controls the better: why does that not count as 'self-control'? Plato may have no real answer, except that Socrates is trying to vindicate ordinary usage, and 'self-control' is a normative term.

Plato does see reflexivity as hopelessly problematic, until it can be explained away by the *Republic* partitioning of the soul.[38]

This evidence is, I think, inconclusive: it is just not clear whether Plato is or should be committed to a meta-agent by worries about the logic of psychic reflexivity. However, there is in any case an alternative way to understand the relevant passages *without* postulating anything over and above the parts of soul. Perhaps Socrates' exhortation in *Republic* IV is addressed to the *rational* part of the soul in particular.[39] This would fit with the way that the rational part is the natural locus of self-control, planning for the overall good, and complex means-ends reasoning—and all those operations in which I reflect on or attempt to reshape my psychological raw materials. Arguably only the rational part is *able* to engage in the kind of long-term self-fashioning that the meta-agent is typically invited to do; so it should come as no surprise if the two turn out to be identical. That Plato invokes the agency of the rational part qua self-fashioner as if it were that of the self as a whole indicates only that reason has a privileged status in the construction of my identity—that my rational part is *me* in a differential degree. [40] I will call this the *asymmetry* claim.[41]

Plato's unargued (indeed unstated) acceptance of the asymmetry claim perhaps stems from the fact that this active role of reason is implicit in any number of everyday, pretheoretical ways of thinking about deliberation and action, without any reification as a genuinely distinct source of agency. The Platonic 'meta-agent' is really just his version of the addressee in 'Pull yourself together!,'

---

38 See Dorion 2007 and 2012. Plato also worries inconclusively in the *Charmides* about whether a 'power' can act on itself: this would presumably be one species of reflexivity, and Plato is here dubious as to its possibility (166a–9c). On the other hand, according to the *Phaedrus* and *Laws* at any rate, what a soul above all *is* is a self-mover, which implies some capacity to be both agent and patient in one.

39 Cf. Irwin 1995, 287–88. The most difficult case for this reading is the passage cited in section I describing the oligarch, in which 'he' arranges his soul so as to *enslave* the rational part. But perhaps even there the 'agent' is really the rational part itself: the idea would be that as the natural ruler of the soul it can only be enslaved with its own active collaboration, when it chooses to internalize and rationalize the ends of its prospective master. *Why* it would ever do so remains a mystery, but there would at least be a nice political parallel: only division within the ruling Guardian class can lead to its overthrow (545c–d).

40 Cf., up to a point, the Kantian reading of Korsgaard 1999.

41 The asymmetry claim may strike some as an obvious Platonic commitment in any case, either on the basis of what Plato says about immortality (on which see my comments in section II above, however) or because he likens the rational part to a human being in *Republic* IX (588b–90e). However, when that passage directs us to cultivate the inner human, Plato does *not* have recourse to the claim that it is my 'true' self, or me in any privileged way.

'Take care of yourself!,' 'How can you live with yourself?' and so forth.[42] The intuitive plausibility of the asymmetry claim suggests that under normal circumstances, it is natural—almost inevitable—for us to identify our cognition disproportionately with that of the rational part. Whether it is stably ruling or ruled, its views and those of the lower parts will run parallel; and what goes on in the rational part will contribute disproportionately to our thinking as we *experience* it, and to its more complex and rationally sophisticated operations above all. (It is at this point that any modern theory would say something about the role of consciousness in the construction of the self.) Reason as ruler deliberates and decides; reason as ruled rationalizes and facilitates. Either way, the rational part, as the locus of discursive reflection, self-control, and decision-making, will be central to how we experience and conceive ourselves as agents. The oligarch, for instance, does not think of himself as a slave to appetite—and with good reason. So far as he can tell, his life has been a model of rationality, dedicated to the calculated, self-controlled pursuit of goals which he reflectively endorses as good. His experience is of a lifetime of restraint and caution (including carefully calculated risk-taking with widows and orphans). Introspection will never tell him otherwise. Only an encounter with Socrates or a reading of the *Republic* has any chance of revealing to him the shocking truth: that he is as much a slave to appetite as the democrats and tyrants he despises.

If the asymmetry claim is our best option for solving the 'meta-agent' problem, is this solution compatible with the analytical reading? The two are in tension, for the latter gives no obvious basis for privileging one part of the soul as more 'me' than another. But they are not outright incompatible. And in fact, the one can help to clarify the other. If the analytical reading is right, then the asymmetry claim holds for a single straightforward reason: not because the

42 Think of the oft-repeated lines from Henley's 'Invictus': "I am the master of my fate/I am the captain of my soul." The captain image is literally Platonic, of course (though spelled out at the level of the city rather than the psyche at *Rep*. 488aff.). And it is so intuitive that readers scarcely notice how odd it is—aren't I also, equally, the second mate of my soul, and the bosun, and for that matter the rigging and the sail? Somehow the identification of the self with the locus of reflexive agency and rational deliberation seems natural, even irresistible. In modern texts this is not only for the reasons cited above, but because of its privileged relation to consciousness. (Cf. the neuropsychologists' idea of the 'interpreter module,' Gazzaniga 2011.) The asymmetry claim thus goes naturally with the idea that only the rational part is connected to consciousness in any systematic way, a finding modern neuropsychology would heartily endorse. I shrink from attributing that further claim to Plato only because, as noted in section 11, I doubt that the *Republic* recognises consciousness as an *explanandum* at all.

rational part is our true eternal self (this is for Plato a separate, eschatological question), or because it is the sole locus of discursive thought (which in the *Republic* at least seems not to be the case), but simply because it can be active in deliberately shaping the whole in a way that the other parts cannot.

The residual difficulty with the asymmetry claim is also simple, and obvious: Plato never states or explains it, even though it would be easy enough for him to do so. It looks as though the problem of the 'meta-agent,' which has so exercised recent interpreters, passes by him quite unnoticed. And that must make it somewhat less likely that an exhaustive scientific analysis of the psyche is really what he has in mind.

V

Here is a third puzzle raised by the analytical reading: call it the *epistemology problem*. On the analytical reading, as I noted in section I, the tripartite theory is *inter alia* a theory of cognition: each part thinks in its own distinctive way. So we would expect Plato to deploy the tripartite theory to explain how different cognitive operations arise from the various parts of soul and their interactions. Yet the tripartite framework is absent not only from early epistemological investigations such as the *Meno* and *Phaedo* (understandably enough), but from the *Theaetetus* and *Sophist*—the post-*Republic* dialogues in which Plato offers his most detailed analysis of how human beings perceive, form beliefs, remember, and make mistakes. Thought is here explained as a kind of silent internal speech, with judgement or belief constituted by inner affirmation or denial (*Theaetetus* 189d–90a, *Sophist* 263e–4a); and our inner 'speaker' is depicted as a monologist, with none of the internal conversation and debate we see depicted in the *Republic* or the *Phaedrus* myth.[43] On the now-orthodox, stylometrically-grounded chronology of Plato's works according to which the *Theaetetus* comes between the *Republic* and *Timaeus*, this is presumably *not* because Plato has come to reject the tripartite theory as false. It must merely be held in abeyance as philosophically unhelpful or irrelevant to the epistemological questions at hand. But on the analytical reading how could *that* be right? The puzzle becomes all the more intense if we note that the *unity* of the soul seems to be critical to its cognition at *Theaetetus* 184d.

Worse, this puzzle is replicated within the *Republic* itself. If the tripartite soul is among other things an analysis of human cognition, it should be

43 Contrast Moline 1978, 13–5, who tries to see the *Theaetetus* and *Sophist* accounts as an expression of the same view.

deployed as such. Instead—and familiarity should not blind us to just how bizarre an authorial strategy this is—having introduced and elaborated his account of the soul as tripartite in Book IV, Plato promptly shelves it without warning for his account of the levels of cognition in Books V–VII. The parts of soul return to the stage for the account of defective souls and constitutions in VIII–IX, the proof that the philosophical life is happiest in Book IX, and the re-banishment of art in Book X. Indeed it is in the discussion of art (including the chunk quoted in section I) that the cognitive dimension of the parts becomes clearest. But this just makes their banishment from the middle Books all the odder. We might try to maintain that at the time of writing the *Meno* and *Phaedo* and again in the *Theaetetus* and *Sophist*, Plato believed that the soul was unitary; but it is highly implausible that he suddenly thought so again while writing *Republic* V–VII, and promptly changed his mind back on reaching Book VIII. The contrast here *must* be a function of his authorial strategy; but what motivates that strategy?

To an epistemologist friend with whom I once raised this question, the answer seemed obvious: 'The rational part is where all the action is!' That is, all the epistemic operations the *Theaetetus* cares about are performed by the rational part alone; so Plato has no reason to allude to the other parts at all in strictly epistemological contexts. Perhaps. But even if this is right about the *Theaetetus*, it seems wrong for the *Republic*. For here and in the *Phaedrus*, the lower parts opine, speak and communicate with each other;[44] so surely the beliefs of the human agent as a whole ought to be understood as constituted by or emerging from that internal dialogue, rather than being referred wholesale to the rational part.[45]

The most straightforward way to solve the epistemology problem would be to deny the explanandum. Perhaps if we look more closely, we can see that the middle Books of the *Republic* actually *do* put the tripartite psychology to epistemological use. I will briefly sketch what seems to me the most promising approach along these lines, and note where it falls short. I will focus on the Cave allegory in particular: since it deals with the transformation of the whole soul by education, this is the most likely place for Plato to have synthesised his epistemology with the tripartite theory.[46] And the most promising place to

44 See Moline 1978, 11–13.

45 As Bobonich 2002 and Lorenz 2006 have argued, Plato may have good reason to change his views about this later on, brought out at *Theaetetus* 184–7. But their reading does not solve the epistemology problem for the *Republic*.

46 C.D.C. Reeve has argued ingeniously for a much fuller and more elaborate 'synthetic' reading of the Cave. It is, however, heavily dependent on his somewhat idiosyncratic reading

find a correspondence between the two is at the lowest level—in the thinking of the prisoners in the Cave.

The Cave depicts the lowest form of cognition, attributed to the bound prisoners, as a taking of shadow-like appearances for realities (515b–c, 520c–e). This seems intended to recall the cognitive state εἰκασία from the Divided Line, which has for its object 'images': shadows, apparitions, and the like, which represent not the Forms (at least not directly) but merely objects in the visible realm (509e1–10a2). The Cave-dwellers' thinking, then, is a kind of 'picture thinking,' organized around images of sensible particulars. The puzzle is to see how this could be anything but a peripheral sort of epistemic condition; and the obvious answer is to add in the argument of Book X (discussed in section I above) that cultural productions—poems, paintings, etc.—are on a metaphysical par with the kinds of image discussed earlier (596a–603b).[47] So people whose thinking about virtue, say, is dominated by 'pictures' taken from Homer and tragedy count as living in an ongoing state of εἰκασία. And in retrospect the Cave can be seen to capture this condition quite nicely. For it depicts them as taking mere reflections for realities—and second-order reflections at that, produced by people unacquainted with Forms. This suggests in turn that more generally εἰκασία includes any cognitive state fixed by *other people's ideas*, conceived as second-order representations, thus including the orator or politician whose conception of justice is not distinguished from 'what most people/the jury/the Assembly consider just.'[48] (Note that it is not just that the Cave-dweller has no way to correct the *content* of other people's conceptions of justice or replace it with something better: he makes a metaphysical mistake as well as an ethical one, assuming that there is nothing else that justice could be.) When the Cave-dwellers identify which image is which, and guess which one is coming next on the basis of past patterns and associations, they are evidently doing what the orator or politician does in predicting what other people will think. To exercise εἰκασία is thus, at the most general level, to think along lines laid down by the culturally enforced conventional wisdom of one's society: to

---

of the Divided Line, and I cannot engage properly with it here (Reeve 1988, 95–100).

47 This is not a new point: see, for example Reeve 1988, 94. Note that this reading has important methodological implications: it means that we do not really have the information we need to make sense of *Republic* VI until we have read *Republic* X, where the Line/Cave account is generalized and integrated with tripartition. Evidently the argument of the *Republic* is only intended to be fully intelligible on a second reading.

48 See Penner 1988 on the impossibility, for Plato, of 'belief-relative sciences.' I mean here to leave it open 'who' the spectators and puppeteers in the Cave are to be understood as in sociological terms, if indeed any particular identification is possible (on which see Wilberding 2004).

believe that courage, for instance, is what Homer represents it as being, including that it is compatible with both hyperaggression and self-pitying lamentation. The characterization makes good sense of the way in which the Cave-dwellers are in a kind of bondage: their thinking is completely controlled by what others present to them, with no independent access to realities unless they can break their cultural chains and turn around.

The important new claim in Book X is that this mistaking of appearances for realities is characteristic of the lower parts of the soul (602c–3a). Cases of optical illusion provide the proof: they show that something in us can form an opinion on the basis of appearances, an opinion resilient against the findings of reason ('encapsulated,' as we now would say). Given the Book IV principle of opposites, that something must be an irrational part of the soul, so this must likewise be what takes art-images as true.

Now a general association of εἰκασία with the irrational parts is at least hinted at earlier on. Plato's post-Cave peroration in Book VII suggests that the cave-dweller is primarily a slave to appetite. The vicious but clever person is highly rational about how to get the things he values; and it is said to be because of feasting, greed and the like—inflammations of the appetitive part—that he comes to be oriented to 'low' objects of cognition (519a7–b5). This somewhat obscure causal claim is confirmed by the account of pleasure in Book IX, which also suggests *how* the ontological mistake characteristic of εἰκασία relates to its ethical defects. Of the appetitively-ruled majority, Socrates says that such people "live with pleasures that are mixed with pains, mere images and shadow-paintings of true pleasures" (586b, compare 'shadow-fighting' at 520c–d). Because of their limited experience, people who are ruled by appetites mistake shadow-pleasure for the real thing, and so take the sources of those pleasures for real goods. The mistaking of an image for the reality—for Plato, the root of all epistemic evil—is the middle term which binds together the ethical errors of the appetitively ruled with the cognitive errors of the shadow-spotters.

So a 'mapping' of Plato's epistemology onto his psychology is available for at least one level of cognition. And this mapping is deliberately hinted at during the middle Books, even if the significance of the hints is unlikely to become clear before Book X. The problem is that to go further, finding a correlate on the Line or in the Cave for the thinking of the spirited part, for instance, quickly becomes fanciful—so much so that I am not even going to attempt it here.[49]

49 A central problem is that there seems to be no one way of thinking characteristic of the spirited part and the people it rules. These must cover a vast range, from the Auxiliary with Guardian potential to the crudest status-obsessed bully. No doubt one could distinguish various spirited types by their different levels of cognitive sophistication and

So it seems that, except for the worst kind of thinking and for the special purposes of Book X, the epistemology and the psychology of the *Republic* are not very well-integrated; and we might take this to show that the tripartite theory is *not* intended as a theory of human cognition after all, as the analytical reading assumes.

What sort of function could the theory serve, then, *without* being a theory of cognition? We might turn for interpretive guidance to the terms in which it is introduced:

> Well then, we are surely compelled to agree that each of us has within himself the same parts and characteristics as the city? Where else would they come from? It would be ridiculous for anyone to think that spiritedness didn't come to be in cities from such individuals as the Thracians, Scythians, and others who live to the north of us who are held to possess spirit, or that the same isn't true of the love of learning, which is mostly associated with our part of the world, or of the love of money, which one might say is conspicuously displayed by the Phoenicians and Egyptians. (435d9–6a3)

The parts are then introduced as explanatory factors behind this distribution of cultural types. Societies get their ethos from the individuals who make them up, and individuals get theirs in turn from their dominant psychological part.

So we might be inclined to propose a reading of the account of the tripartite soul as strictly a theory of *personality*, and by extension of culture. And this is, after all, the realm in which it is put to its most extended use, in the analyses of defective constitutions in Books VIII and IX. This part of the *Republic* begins with a reaffirmation and reminder of the Book IV passage just cited:

> And do you realize that of necessity there are as many forms of human character as there are of constitutions? Or do you think that constitutions are born 'from oak or rock' and not from the characters of the people who live in the cities governed by them, which tip the scales, so to speak, and drag the rest along with them? (544d5–e2)

However, as a way to evade the epistemology problem, this cannot work. For the tripartite account is after all presented as *explanatory*—not merely as a typology. On almost any reading the parts of soul are causes of behavior, standing internal forces which shape our agency; and their cognitive functioning is

---

insight, but it is hard to see how these would correspond to the different levels of the Line or Cave.

surely essential to that explanatory role. Even if the desires of the lower parts are to some extent 'good-independent,' they must be shaped by the perceptual and conceptual resources of those same parts: that is, given the possibility of psychic conflict, their desires cannot be too blind to produce action. When Leontius experiences an impulse to look at corpses, whatever perception triggers his anticipation of pleasure in doing so cannot be coming from anywhere but the appetitive part itself—neither the spirited nor the rational part will register those corpses under any arousing or attractive description. Likewise in the *Phaedrus* myth, the dark horse perceives and responds to the beauty of the beloved all by himself. The parts admittedly exploit each other's capacities: in particular, appetite would love to bend the rational part's superior powers of instrumental reasoning to its own ends. But even that aim presupposes that each part has sufficient mental powers to conceive those ends and feel their attraction.[50]

All this is just to reaffirm the obvious: personality is the behavioral expression of cognition, and tripartition cannot explain how we behave unless it explains how we think. So the epistemology problem remains a puzzle. All we can say is that Plato seems to alternate deliberately between two very different explanatory perspectives on human cognition. When his concern is with agency, character and moral psychology, Plato's perspective is psychodynamic, and focusses on distinguishing the subpersonal agents or subsystems which can explain mental conflict, virtue and vice, and differentiation of character: his favored vehicle for this kind of project is the theory of the tripartite soul. But when his concerns are more narrowly epistemological (as in the *Meno*, *Phaedo*, *Theaetetus*, *Sophist*, etc.), this kind of explanation drops out in favor of the hierarchical distinction-drawing of the Line and Cave, and the post-*Republic* analysis of concepts like knowledge and judgement. For reasons which are to me unclear, only in Book X of the *Republic* do the two perspectives explicitly come together, for the rather specialized purpose of explaining our cognitive-emotional reactions to art.

50 Note also that Plato generally prefers to refer to the 'parts' of soul as 'forms' or 'species' [εἴδη], which brings out that the rational, spirited and appetitive parts are three species of the genus ψυχή. And according to the *Phaedrus*, the defining characteristic of that genus is self-motion (245c–e). So the 'parts' of soul are really three different types of self-mover (hence the appropriateness of Plato's animal imagery). That means that each must have powers adequate to *independently* get action off the ground; which in turn implies sufficient cognitive resources to grasp states of the world and fix on ends of action.

VI

Analytically read, and taken at a somewhat high level of abstraction and generality, Plato's tripartite soul has a distinct family resemblance to the embodied psyche studied by contemporary neuropsychology. But whether Plato's theory *should* be analytically read is another question, and I have discussed three objections to doing so. The first, the 'revised *homunculus*' problem, was that—in contrast to contemporary scientific theories—the tripartite account does not seem to aspire to analyze basic mental operations like believing, desiring or deciding, but merely gives 'locationist' explanations of them. Second was the *meta-agent* problem, namely that (again in contrast to contemporary scientific analyses), Plato's analysis does not seem intended as exhaustive: for he wavers on the identity of the soul so analyzed and the self, at times postulating an agent above and beyond the parts. Third was the *epistemology* problem. For the most part, Plato does not actually deploy the tripartite theory as a theory of cognition; and it is not easy to see how his epistemology could be grafted on to it. But this strongly suggests that it is *not* intended as a full scientific analysis of the psyche, for this must be a theory of cognition before it can be anything else. I have offered the best responses I can to these challenges, but in each case the results are, in my view, inconclusive. The upshot is that it remains an open question just what *kind* of theory the theory of the tripartite soul is supposed to be. Is the analytical reading really right to take it as the kind of analysis familiar to us from contemporary natural science? And if it is not that kind of theory, what is it?