Carrie Figdor
Department of Philosophy
University of Iowa

Original Scientific Paper
UDK 165.194:159.954/.956
165.3/.6

THE FALLACY OF THE HOMUNCULAR FALLACY[1]

Abstract. *A leading theoretical framework for naturalistic explanation of mind holds that we explain the mind by positing progressively "stupider" capacities ("homunculi") until the mind is "discharged" by means of capacities that are not intelligent at all. The so-called homuncular fallacy involves violating this procedure by positing the same capacities at subpersonal levels. I argue that the homuncular fallacy is not a fallacy, and that modern-day homunculi are idle posits. I propose an alternative view of what naturalism requires that reflects how the cognitive sciences are actually integrating mind and matter.*

Keywords: *homuncular functionalism, homuncular fallacy, mechanistic explanation, psychological models, naturalizing the mind, psychological explanation*

1 Introduction

The most familiar, and arguably received, theoretical framework for an adequate naturalistic explanation of mind is homuncular functionalism (Lycan 1991; Dennett 1975; Fodor 1986; Cummins 1983).[2] The homuncular functionalist strategy is to explain a cognitive capacity of a whole in terms of the organized not-quite-cognitive operations of its parts. Because of its part-whole decompositional nature, homuncular functionalism can be seen as a special case of mechanistic explanation, a leading contemporary view of scientific explanation (e.g., Machamer et al. 2000; Bechtel 2005; Glennan 2002; Craver and Tabery 2015).[3] Unfortunately, homuncular functionalism cannot

1 This article builds on a position initially articulated and defended in Ch. 8 ("Literalism and Mechanistic Explanation") of Figdor (2018). The book offers a comprehensive discussion of the problem of interpreting psychological language throughout biology.

2 Throughout, I will use "cognitive", "psychological" and "mental" interchangeably, and "capacities", "functions", "operations", and "activities" interchangeably to denote what entities are able to do and what they actually do on occasion. Nothing here turns on the difference between ascribing to an entity an ability to do X or to perform a function X and describing an entity as doing X or performing the function X.

3 I will discuss homuncular functionalism's relation to ongoing debates in psychological explanation in more detail in section 3. In brief, however, homuncular functionalism counts as a type of functional analysis, but its assignment of the homunculi to components makes the framework mechanistic (Craver 2001; Piccinini and Craver 2011)

 doi: 10.5937/BPA1831041F

bask in the popularity of the new mechanistic philosophy. Contemporary mechanism, alongside other developments, shows that modern-day homunculi are idle posits, just as their Aristotelian namesakes once were to historical mechanists. Even further, the mind is being naturalized using explanatory tools that do not require them.

In what follows I first present the mechanistic explanatory framework defended in recent years and its relation to homuncular functionalism. I then show how accepted mechanistic explanations regularly violate the characteristic constraints of homuncular functionalism without triggering the epistemic disaster that motivates the framework. In addition, there is no explanatory justification for a psychological exception to the rule. As a result, homunculi are bogeymen in principle and are treated as such in scientific practice. I conclude by suggesting an alternative view of what naturalism requires in the context of contemporary efforts to integrate psychology and neuroscience.

2 Homuncular Functionalism and Mechanistic Explanation

The term "homunculus" was originally introduced to label the preformed little men posited within the Aristotelean explanatory tradition to explain how the human adult could emerge from an embryo (Maienschein 2017). Homunculi resolved the problem by holding that everything in the adult is already in the embryo – not in the sense accepted today that the adult *develops* from what is already in the embryo (ignoring the role of environmental factors), but in the sense that the embryo already contains the final result in miniature. A kernel of this view persists today in the claim that an embryo (or even a zygote) is an unborn child with rights.

Seventeenth-century supporters of the then-new mechanistic approach to explaining nature rejected Aristotelian homunculi as idle posits. In general terms, both historically and contemporaneously, a mechanistic explanation is an explanation of a capacity of an entity in terms of that entity's constituent entities and activities and their organized causal interactions (e.g., Machamer, Darden, and Craver 2000, Glennan 2002, Bechtel and Abrahamsen 2005). Moliere's ridiculing reference in *Le Malade Imaginaire* to the dormitive virtue of a sleeping potion was of a piece with this rejection of homunculi by mechanists. Fodor (1968: 627) recapitulates this ridicule for contemporary readers with a farcical computational account of how to tie one's shoes in terms of a little man inside who follows a rulebook for tying one's shoes.

Historical mechanists also understood mechanisms in a narrow way that reflected the machines with which they were familiar, such as clocks or

even if psychological explanation in general may not be (Weiskopf 2011). According to functionalism in philosophy of mind, which I assume here for the sake of argument, mental states are individuated by their functional roles in a cognitive system.

mills. This narrow conception informed La Mettrie's (1748/2017) defense of mechanistic explanation of mind, as well as Leibniz's rejection of it. To Leibniz (1714/1989), "perception, and what depends on it, is inexplicable in terms of mechanical reasons, that is, through shapes and motions": one could imagine walking into enlarged version of a machine structured to make it think and perceive, and inside "we would only find parts that push one another, and we will never find anything to explain a perception." Homuncular functionalists side with La Mettrie in terms of the possibility of a mechanistic explanation of mind, but the lingering question has been how to bridge the explanatory gap that Leibniz and others identified.

To do this, homuncular functionalism imposes three characteristic adequacy conditions on its decompositional explanations of cognitive capacities. First and second, the subcapacities ascribed to the parts must be distinct from and simpler than the cognitive capacity of the whole that they explain. Lycan explains both conditions as follows:

> [H]omunculi can after all be useful posits, so long as their appointed functions do not simply parrot the intelligent capacities being explained. ... We account for the subject's intelligent activity, not by idly positing a single homunculus within that subject whose job it simply is to perform that activity, but by reference to a collaborative team of homunculi, whose members are individually more specialized and less talented. (Lycan 1991: 260)

The homunculi are also variously described as less problematic, elementary, primitive, or less clever (Cummins 1983; Dennett op.cit.). These two features give rise to their being called homunculi, although the label is a bit misleading: the original homunculi were exactly similar in all relevant respects to what they were miniatures of, whereas modern-day homunculi are explicitly barred from being exactly similar in these respects.

Third, the call for less demanding subfunctions must be iterated at each level of the decompositional explanation until the subfunctions are no longer cognitive and the homunculi are "discharged" (Dennett 1975). As Fodor (1968: 629) puts it, a modern-day homunculus is a "representative *pro tem*" for a system of instructions "that makes no reference to unanalyzed psychological processes". The iteration of subfunctions continues until the activities of the smallest parts within the explanation's scope are entirely devoid of intelligence. Further decompositional explanation – to the levels of cellular mechanisms, for example – proceeds by means of capacities that are entirely non-cognitive (i.e., physical). This gradualism enables modern-day homunculi escape Molierian/Fodorian ridicule; it is the cognitive analogue of the embryo-to-adult relationship as now understood. Since we already accept a naturalistic theory of how an embryo can develop into an adult, it seems but a small step to understand a framework that posits a similar relation

between physical and psychological capacities, at least synchronically (and maybe diachronically).

These three adequacy conditions are motivated by the epistemic worry that ascribing the psychological capacity of a whole to a part is non-explanatory or idle, as the uselessness of the original homunculi in a mechanistic framework appeared to demonstrate. Bechtel (2009: 561) writes that "assuming a homunculus with the same capacities as the agent in which it is posited to reside clearly produces no explanatory gain", while Dennett (1975: 171) adds that to avoid being "question-begging" the most fundamental posits within the explanation must not be supposed to perform tasks or follow procedures requiring intelligence.

The homuncular fallacy involves violating these requirements for distinct, ever-stupider capacities in a decompositional series that gradually gives way to non-cognitive capacities. To commit the fallacy is to fail to naturalize the mind.

3 Troubles for Homuncular Functionalism

The homuncular functionalist's adequacy conditions have not gone unchallenged. For example, Margolis (1980) argues that there is no *a priori* reason to restrict explaining capacities to "stupider" ones. Freudian psychology posited a complex unconscious which explained simple or complex patterns of behavior. Although the Freudian explanatory framework may have been rejected by many, it was not rejected because it violated the second adequacy condition.

I will challenge homuncular functionalism instead by undermining the epistemic worry that motivates positing homunculi in the first place. The contemporary explanatory practices that motivate the new mechanism also show that the adequacy conditions of homuncular functionalism are unnecessary.

First, contemporary mechanists do not restrict mechanistic explanation to "exclusively mechanical (push-pull) systems" (Machamer, Darden, and Craver 2000: 2). For example, restriction to push-pull mechanisms "makes the concept of a mechanism too narrow to accommodate the diverse kinds of mechanism in contemporary neuroscience" (Craver 2007: 4). Because their framework is intended to capture actual scientific explanatory practices, contemporary mechanists emphasize a characteristic explanatory method rather than characteristic kinds of activities. Margolis' criticism of homuncular functionalism is, in effect, a special case of this point.

Second, these same explanatory practices regularly involve ascribing the same capacities to wholes and their parts without generating epistemic distress. The same kinds of activities appear at different spatiotemporal scales and different decompositional levels, and their contributions to these explanations are not idle. A piece of machinery lifts boxes because its engine

has a camshaft that lifts valve covers. If a piano string is vibrating (exhibiting simple harmonic motion), this will be in part because the molecules in the string also vibrate. A monkey learns to swing through the trees by grabbing and releasing vines, and presynaptic neurons in the monkey's hippocampus release glutamate in the process of long-term potentiation, theorized to be a mechanism of learning. Sober (1982:421) raises this point when he notes that the apparent "emptiness" of homuncular explanations does not stem from ascriptions of the same operations to parts and wholes, but from a shift from tokens to types. To borrow his example, no epistemic error is generated when we hold that planets rotate in part because the nuclei of their atoms rotate. In a similar vein, Cartwright (1983: 145) notes the repeated use of the harmonic oscillator model at multiple scales. To the extent that such repetition is problematic, it is because we are no closer to an explanation of (e.g.) rotating as a type of activity ("what unites planets and atoms"). But that's not a reason to deny that planets and their atomic nuclei both rotate.

It follows from these two points that cognitive capacities are just as apt as any other type of capacity for being ascribed to both wholes and their parts, at least in principle.

Third, there is actual disregard of the first and second homuncular functionalist restrictions in scientific psychological practice. For example, Sutton and Barto (1981) theorized about adaptive elements that would comprise adaptive systems, intending their temporal difference (TD) model of reinforcement learning to apply to both. The model ascribes such capacities as anticipating and predicting to the adaptive systems that exhibit this form of learning. Although the model was developed based on animal learning data (including humans), they explicitly suggest neural assemblies as possible targets of their model in addition to humans, dogs, and other adaptive systems with which we are more familiar. Two decades later, Suri and Schultz (2001) implemented the TD model in a connectionist simulation of actual neural activation patterns, showing that Sutton and Barto's speculative suggestion was not idle. Similarly, albeit in rhetorical fashion, Wimsatt (2006: 461) remarks: "Memory – a property of molecules, neural circuits, tracts, hemispheres, brains-in-vats, embodied socialized enculturated beings, or institutions?" It is unmotivated to prohibit *a priori* the ascription of memory to both wholes and their parts, given that it is in fact being so ascribed by scientists across many fields or is clearly considered an empirical possibility that can help explain human and nonhuman behavior rather than provoke explanatory failure.

The homuncular functionalist might respond to such examples by saying that (e.g.) real memory is only being ascribed at the personal or whole level. But this response risks trivializing the view: whenever a cognitive capacity is apparently being ascribed at a subpersonal level, the homuncular functionalist can always dismiss it as not real. As a naturalistic theory, homuncular functionalism should

not take for granted definitions of cognitive capacities, or interpretations of cognitive ascriptions, that entail that the theory cannot be falsified by evidence from the relevant sciences. Yet while the homuncular functionalist must posit such conceptual changes whenever the same word-forms are used at both personal and subpersonal levels, her reinterpretation of their meaning cannot be the default interpretation given that multi-level ascriptions of capacities in mechanistic explanations are often assumed to be conceptually continuous. For example, "rotates" picks out the same capacity exhibited by planets, ballerinas, and atomic nuclei even if each rotates in its own way.

But what then of Bennett and Hacker's (2003) mereological fallacy? This is the fallacy of using psychological terms to ascribe psychological capacities to parts when the very meaning of these terms makes such ascriptions nonsensical (according to Hacker's Wittgensteinian logico-grammatical orthodoxy). Bennett and Hacker explicitly argue that it is a *conceptual* mistake to ascribe (fully) cognitive capacities to parts. Dennett himself (Bennett et al. 2007: 88–89) responds to this fallacy on behalf of the homuncular functionalist: when the terms are used to ascribe capacities to parts, they ascribe "hemi-demi-semi-proto-quasi-pseudo" cognitive capacities (that is, homunculi) to the parts. As a result, no mereological fallacy is committed.

Dennett's response illustrates the point made above. He makes explicit the homuncular functionalist's need to posit conceptual change to explain away psychological ascriptions at subpersonal levels. Escape from the mereological fallacy comes at the cost of imposing *a priori* conceptual constraints of its own. (If it were a theory of meaning, one might say this *just is* the theory.) The problem remains that no such conceptual change can be taken for granted in the light of general scientific ascriptive and explanatory practice. Even if (*pace* Dennett) the psychological terms *are* being used to ascribe the same cognitive capacities to the parts, we have no reason to think that doing so runs afoul of any epistemic worries. Absent this epistemic motivation, the straightforward response to the mereological fallacy is to argue that psychological predicates are not conceptually barred from being ascribed sensically to the parts.[4]

This yields a puzzle. The ascription of the same capacities of wholes to their parts within a mechanistic explanation is not epistemically problematic in general. To the contrary, it is par for the course. Nor does scientific psychology appear to care about the homuncular functionalist constraints. Perhaps reprising *objects* at the level of parts really is epistemically idle: maybe one can't, on pain of epistemic idleness, build a dog using tiny dogs, rather than cells, as parts. But the epistemic idleness of this repetition for objects, even if true, clearly does not extend to capacities, not in general science and not in contemporary psychology.

An obvious solution to the puzzle is to show that psychological capacities are exceptions to the rule for adequate mechanistic explanations. Maybe there

4 In Figdor (2018), I argue at length against their view and against the mereological fallacy. For present purposes, the homuncular functionalist's response is all that matters.

is something about *the mind* that makes repetition of *psychological* capacities at the level of parts invariably explanatorily pernicious.

This defense of homuncular functionalism might start from the idea that repeating a capacity of the whole at the level of parts would leave the explanation incomplete. But while incomplete explanations are relatively undesirable, they are not circular, idle, or question-begging merely because they are incomplete. The homuncular functionalist might add that in this case the mystery of the mind at the level of the whole would simply be repeated at the levels of the parts. Since nothing will have been done to dispel this mystery, the purported explanation would be viciously idle and circular, as feared. In incomplete non-psychological mechanistic explanations, in contrast, we know in principle how to fill in the gaps even when the same capacities are ascribed to parts. It's the fact that *the mind itself* is *especially* mysterious that yields a special problem.

The response implicitly concedes that if we understood better what psychological capacities are, it wouldn't actually matter if they were ascribed to parts. After all, the ban on repetition is intended to address a worry about a lack of explanatory gain. Non-psychological mechanistic explanations reveal that lack of explanatory gain is distinct from repetition. Repetition provides explanatory gain in that a repeated operation or capacity ascription still fills in details of mechanistic explanations, but the gain it provides will remain incomplete without illumination of the capacity. So the basic, and very real, issue is how to illuminate cognitive capacities. Homunculi are posited specifically to provide that illumination, of course, but their explanatory contribution depends on their gradually increasing stupidity (or gradually decreasing intelligence) as the decomposition proceeds, and so is not independent of the issue of repetition. One might wonder how much illumination is actually provided by the homuncular functionalist metaphors for this process. Setting that issue aside, if we can illuminate the mind naturalistically yet independently of the issue of repetition, we will have a naturalistic framework that does not require either homunculi or making psychology an outlier among the sciences.

Consider why at least some non-mental capacities are non-mysterious and can contribute to mechanistic explanations by being ascribed to parts as well as wholes. For example, why is rotating not mysterious at any level at which it is ascribed? One might provide two complementary mystery-dispelling reasons. Perhaps there are others, although these appear to suffice. First, we have some prior understanding of what rotating is – we've played with spinning tops, watched ballerinas, operated drills, and so on. So when entities that are too far away, too big, or too small to observe unassisted are said to rotate, the ascribed activity is not wholly mysterious, whatever it is ascribed to. Call this the familiarity condition.

Second, we have equations of angular momentum that help us distinguish instances of rotating from non-instances, even if the items doing the rotating

are as distinct and as complex in their own ways as planets, ballerinas, and atomic nuclei. We individuate rotating as a type using a scientifically accepted method for distinguishing different kinds of motion, and we use that method to pick out tokens of rotating independently of our parochial perspective on which things rotate. The way each item achieves or exhibits its capacity to rotate may differ, but that's not problematic. The equations help guide our ascriptions by providing a standard for determining when things we think are rotating really are and when they aren't. Call this the objective constraints condition.

In sum, armed with a scientifically accepted kind of evidence for and constraints on the ascription of rotating to an entity, plus some prior understanding of that capacity, there is no fear of perpetual mystery that might motivate a ban on ascribing rotating to parts within mechanistic explanations of the rotating of wholes.

In the case of psychological capacities, the familiarity condition is clearly met. We do have some familiarity with what these capacities are from our own case. There's nothing wrong with that; we need to start explaining the mind from somewhere. We don't also need to think we're infallible or that the mind is transparent to introspection.

What is new and significant is that the objective constraints condition is also being met. The contours of this type of illumination in psychology are now discernable through the use of mathematical models of cognitive capacities. The mathematical models I have in mind are those expressed using equations that formally describe empirical relationships the way mathematical models of non-psychological phenomena (such as the Hodgkin-Huxley model of the action potential) do. A psychological example of such a model is the drift-diffusion model (DDM) of two-choice decision-making, first proposed by Roger Ratcliff (1978) and subsequently elaborated, tested, and extended by Ratcliff and colleagues and other psychologists. (The TD model mentioned above is another.) This model proposes cognitive processes of evidence accumulation and assessment to a threshold, at which point a decision is made and a behavioral response is given. For example, subjects (often undergraduates) may be asked to press a key to report whether an image is of a house or a face, and the experimenter manipulates the clarity of the image. The model posits cognitive processes that mediate between the stimulus-response relationship in a way that captures the speed-accuracy tradeoff: the relation between the clarity of the evidence and the speed and accuracy of the subsequent responses. Given the same degree of noisiness in the stimuli, subjects make more mistakes (relative to benchmarks) when they are instructed to respond quickly and respond more slowly (ditto) when instructed to emphasize accuracy.

What is formalized by such models are the observable behavioral patterns of people (and some nonhuman species) from which we infer to a cognitive capacity or combination of them that can yield these patterns. This is crucial

when we are interested in the capacities of entities that are not already in the cognitive club by hypothesis or general consent. For example, the DDM equations provide a standard scientific means to identify tokens of the posited decision-making processes independently of our parochial understanding of mind. In the case of undergraduates, the appropriateness of ascribing internal cognitive processes of accumulating and assessing evidence to a decision threshold is assumed. Human behavioral data was used to develop the model in the first place. But the DDM has also been used to examine whether fruit flies' decision-making is affected by a genetic flaw that also affects humans (Dasgupta et al. 2014). It was not given that the model could be used for fruit flies. It was a matter of empirical test. This fact grounds its objectivity: it is not up to us to decide where it may fit. We may not otherwise have evidence of a behavioral regularity or of its similarity to the human behavior on which we base cognitive ascriptions. Ordinary observation can even impede this recognition. For this reason, the model can be used to determine when something is behaving relevantly similarly to the way we do when we have made a simple decision and then act on it, even if the entity is not of a kind that we intuitively think of as being able to make decisions.

Of course it does not *follow* that the natural language expressions used to interpret the equations are being used in the same way when the models are extended successfully in new domains. But extension by means of formal models is part and parcel of the same scientific methodology by which we ascribe capacities such as rotating or oscillating across vastly different domains as well. The practice increasingly includes cognitive science and social science (e.g., Irvine 2016, Froese et al. 2014). Moreover, in contemporary network science, the use of the same models at multiple spatiotemporal scales is integrated into the practice of explaining the behavior of wholes in terms of the behavior of their parts (e.g. Alon 2007, Baronchelli et al. 2013). This is illumination of the sort Newtonian mechanics achieved between the terrestrial and celestial domains via one set of laws of motion that applied to both. Not incidentally, Newton's laws naturalized the celestial realm, whereas before it was mysterious and divine.

The fact that we know how to satisfy the objective constraints condition for psychological capacity ascriptions does not imply that we will not find other types of objective evidence. Cognitive neuroscientists are actively seeking neural activation patterns that will enable us to "mind-read" by using neural behavior as the basis for our inferences (Poldrack 2006; Roskies 2014). Such evidence could complement behavioral evidence, and in cases of conflict the outcome for ascriptions will depend on other factors. It also does not imply that a single model for each intuitively individuated cognitive capacity will suffice. Naturalizing the mind will require confronting at least three distinct sources of complexity in traditional psychological ascriptions: distinguishing and relating various types and subtypes of cognitive processes; distinguishing the non-epistemic and epistemic goals of our traditional cognitive-ascriptive practices; and articulating differences in the grounds

and types of meaning adjustment as word-forms are used in new domains. These factors will have to be disentangled as naturalization proceeds. The present point is that mathematical models show how we can begin to satisfy the objective constraints condition for the mind. If we have evidential means to ground cognitive ascriptions to parts, and the ascriptions are being made on these grounds, and the evidential means and ascriptive practices follow standard scientific canons, and all of this violates the homuncular functionalist constraints on an adequate naturalistic explanation of mind, *tant pis* for homuncular functionalism.

Dennett describes discharging as the point at which there are no questions about intelligence being begged. Cognitive capacities are no doubt more complex than rotating. But what would be missing if at every level at which a cognitive capacity were ascribed on the basis of a model, we could then explain mechanistically how the entities at that level realized that capacity? The explanation would be incomplete until we filled in these details, but there is no reason to think we have perpetuated mental mystery. Naturalization without homunculi implies that psychological capacities may not be eliminable. But explanation does not require elimination. It requires eliminating mystery. Only if we think the psychological is essentially mysterious does it follow that eliminating the mystery of the psychological requires eliminating the psychological. Oddly enough, the homuncular functionalist method implies that psychological concepts and the capacities they pick out are not just mysterious now, but essentially so.

Note that the phrase "psychological model" (or "cognitive model") has a longer history and wider scope that includes more than just the mathematical models discussed above. The phrase also encompasses boxologies or other informal functional analyses of cognitive processing, as well as computational models of cognition. The latter are connectionist (or hybrid symbolic-connectionist) networks used as stand-ins for neural networks and their activation patterns; they are models of possible realizations of cognitive capacities by neuron-possessing creatures that by hypothesis already belong to the cognitive club. Note that these networks may also be called mathematical models – I am not seizing the label but merely using it in this paper to be precise about the sort of psychological model that can play the evidential role for psychological ascriptions that (e.g.) equations of angular momentum do for ascriptions of specific motions.

The important difference between these types of models and the mathematical models of interest here is that one cannot use these other types of models on their own to guide and constrain psychological ascriptions across domains. They may satisfy other explanatory purposes, but not the purpose of determining which things have cognitive capacities. Boxologies do not include operationalized behavioral signatures of the posited capacities, although they can be augmented with them. If formal, this would supplement the boxologies with the sort of evidence that mathematical models contribute; if the behavioral

patterns are not formalized, the augmentation would fall short of providing objective constraints on psychological ascriptions to non-human domains.

Artificial neural networks do link input with output, but the operationalization consists of assigning numerical values (vectors) to whatever the inputs and outputs happen to be. The evidential work of interest here is done by identifying the stimuli-response pairs that the input and output vectors represent. Again, if these relationships are not formalized, we would not have an objective means to extend the network to new domains, even if we set aside the fact that cognitive capacities in entities without brains (such as plants or slime molds) are beyond the scope of these models from the start. In addition, connectionist networks are used widely outside psychology. Whether what they represent is a psychological process at all is not determined by anything intrinsic to the network. If or when future artificial network designs make such identification possible, they would be an additional tool for investigating internal evidence of cognition, alongside the "mind-reading" research mentioned above.

In sum, we now have the tools to overcome the evidential gap for ascribing psychological ascriptions to nonhumans in an objective manner. Mathematical models of cognitive capacities rely on our normal understanding of the capacities being modeled and so satisfy the familiarity condition for illumination. When patterns of behavior of humans is captured formally, we can see if behavioral data from nonhumans satisfies the model, whatever we may think of those nonhumans or their behavior. Formalization provides clear constraints on when we are entitled to ascribe to an entity the capacities that we ascribe to humans using psychological language. This procedure satisfies the objective constraints condition of illumination. Ascriptions of harmonic oscillation or rotation to entities at multiple scales follow the same procedure. In the case of the mind, unlike those cases, the ascriptions are made by inference from the observed patterns. But that is true of our ascriptions of cognitive capacities to each other. Finally, these possible extensions include entities that are parts of wholes to which we also ascribe the capacities on the basis of the behavior captured by the same model.

Armed with this alternative, we simply do not need homunculi to naturalize the mind. The dormitive virtues lost their explanatory power because virtues no longer had explanatory force within the mechanistic explanatory framework. Modern-day homunculi can go the way of the dormitive virtues. In contemporary scientific psychology, they too are now idle, along with the discharging requirement associated with them.

4 Model-based Naturalization, Mechanisms, and Autonomy

In the last section I distinguished among types of psychological models in a way that reflects the present paper's concern about the viability of homuncular functionalism give contemporary explanatory practices in the cognitive sciences. In this section I will consider how the model-based framework for naturalization intersects with two current debates regarding scientific

explanation in general and psychological explanation in particular. Their common starting point is mechanistic explanation. First, do laws or models provide explanations at all? Second, is psychological explanation autonomous from neuroscientific explanation? I will show how the view is neutral regarding these debates. The questions they raise remain outstanding even if we reject homuncular functionalism in favor of model-based naturalization.

First, do models or laws *explain*? This is a long-running debate between the Hempelian covering-law model of explanation and mechanistic explanation as the two basic conceptions of scientific explanation. For example, Craver 2006 calls the Hodgkin-Huxley equations "phenomenal models" that can be used for prediction, like covering laws, but they do not *explain* because they do not detail the mechanisms. Presumably the DDM and other mathematical models of cognition also count as phenomenal models and so would also not explain for the same reason. More recently, mechanists have proposed that psychological models are mechanism sketches, which are incomplete mechanistic explanations or when filled in with the details turn into mechanistic explanations (Piccinini and Craver 2012). This position accepts that models play a role in scientific explanations, but ties their ability to explain to that of mechanisms. Opponents claim that laws or phenomenal models provide explanations independently of whatever explanatory work is provided by detailing the mechanisms (e.g. Batterman and Rice 2014, Chirimuuta 2014).

Relative to this debate, it is sufficient for my purposes to claim merely that laws and models contribute to explanation, for the feature of interest here is their applicability to multiple levels of mechanisms. Phenomenally adequate cognitive models combine evidence of patterns of behavior with a familiar psychological conceptual framework that provides some understanding of the capacities that may be ascribed at multiple levels. It is a further question whether the models' contribution at any level to which they apply is due to their own explanatory power or because they function as mechanism sketches. They certainly do provide constraints on capacity ascriptions, non-cognitive or cognitive. Having such objective individuation criteria is an important advance in understanding, whether we want to consider this advance an explanation or not. Models and mechanisms may simply have a symbiotic relationship. As Sober (1982: 421–422) remarks with regard to laws, if we are told an organism digests in part because parasites in it digest, "we now want to know what laws govern the way organisms obtain energy from their environments, and how those laws apply simultaneously to hosts and the parasites they house."

Homuncular functionalism, in contrast, occupies an awkward position in relation to this debate. It holds that an activity of a part does no explanatory work if it is the same type of activity as the activity of the whole to which the part belongs and which the part-level activity is supposed to help explain. But it also holds that an activity of a part *can* do explanatory work if it is related to the capacity of the whole as being lesser on a continuum of similarity that gets whittled down to nothing. In other words, what makes a homunculus

dissimilar makes it explanatory and what makes it similar perpetuates explanatory idleness. Neither law-based nor mechanistic explanation affirms these claims. For example, the equations of angular momentum unify a planet's rotating and an atomic nucleus' rotating. The laws reveal an important similarity across domains, and for some philosophers (and scientists) this unity is the essence of explanation. For mechanists, it simply doesn't matter if planets and their atomic nuclei both rotate, for the explanation is provided by showing *how* they rotate, not that the mechanisms are relevantly similar or different. An adequate answer to the 'how' question neither requires nor rules out multiple realizability of the explanandum phenomenon.

Second, are psychological explanations in some important sense autonomous from neuroscientific ones? On the assumption that the latter are mechanistic, this question becomes a special case of the first debate, although the question of the autonomy of psychology from neuroscience predates the rise of the new mechanism. In contemporary terms, the claim that cognitive models are mechanism sketches (Piccinini and Craver op.cit.) or that they need to be mapped to mechanisms to have explanatory power (Kaplan and Craver 2011) is a way to argue that psychological explanations are not autonomous. In contrast, Weiskopf (2011: 322–23) distinguishes between psychological models that aim to explain psychological phenomena in semantic, intentional, or representational terms, and those that aim to explain psychological phenomena in non-psychological (e.g., neurobiological) terms. The first group includes connectionist networks whose states, while subpersonal, are interpreted in representational terms. Focusing on the first group, Weiskopf argues that models in this group are not mechanistic because their components don't correspond to the parts of the modeled system in any straightforward way. This would be a way to defend the autonomy of psychological explanation.

The model-based view of naturalization is neutral regarding the autonomy debate. It allows for psychological ascriptions at personal and subpersonal levels. Whether these models are mechanistic, explanatory on their own, or autonomous from neuroscientific models or mechanisms are further issues – the naturalization project does not depend on how they are resolved. Homuncular functionalism, in contrast, implies that ultimately psychology can't be autonomous. The view straddles Weiskopf's categories: the explanation starts out in the first category but ends up in the second. So even if psychology starts out autonomous, the upshot of the explanatory framework is to undermine that autonomy step by step even if that is not its goal.

As a final issue, model-based naturalization suggests that the traditional debate between reductive vs. non-reductive physicalism must be reconfigured for a scientific context in which the same formal structures are employed at multiple scales. An implicit assumption of the traditional debate is that these are distinct conceptual schemes whose distinctness is in part tied to particular levels. The gradualist framework of homunculi takes on this

implicit assumption. The model-based view of naturalization makes reduction a levels-relative (and not just capacity-relative) affair. If psychology reduces to neuroscience at one level, this result cannot automatically be generalized to all levels. Moreover, if assigning entities to levels is difficult, assigning capacities to levels will also be difficult (Stinson 2016). Even the most favorable cases of reduction may require significant simplification of the phenomena.

5 Concluding Remarks: Naturalization Without Homunculi

The contemporary explanatory context for psychology is one in which models of capacities are apt in principle for use at any level in a mechanistic hierarchy. If the models apply to wholes and their parts, so be it. Positing ever-stupider homunculi in this explanatory context is like insisting that only ballerinas really rotate even though the same equations of angular momentum apply to their atoms. We don't need stupider capacities at each level. To the contrary, therein lies the perpetuation of mystery.

It is difficult to identify what homuncular functionalism gets right, if anything. Perhaps it is just the idea (surely not unique to homuncular functionalism) that there is a special epistemic problem involved in naturalizing the mind. But homuncular functionalism did not identify the problem correctly. The unique challenge faced in psychology is the problem of distilling from our homegrown, first-personal, often introspective understanding of the psychological those features that are specific to humans but contingent to possession of the capacity. Until recently, we had no idea how to gain a non-anthropocentric perspective on the mind. Now we do. Models can help us distinguish those aspects of a cognitive capacity that are matters of human realization and those which are arguably what it is to have that capacity. Where subpersonal entities differ from humans (or other wholes with minds) need not make any difference to their possession of the capacity, just as the fact that planets and atomic nuclei lack arms and legs does not deprive them of the capacity to rotate.

Have I argued for a different type of homunculus? No – the proposal is not that neurons are little men. The proposal is that parts of men can, in principle, do what men do, and that we have the empirical tools to discover whether they do or not. Have I argued that consciousness can be explained with non-intrinsic properties? No – I don't know how consciousness will end up being explained. Have I shown that the mind can be explained naturalistically? No – I've taken for granted here that it can be. What I have shown is this: if the mind can be explained naturalistically, we have an alternative to the homuncular functionalist framework that is a natural fit with the way many accepted explanations in the sciences are actually formulated, and with the way contemporary cognitive researchers are pursuing their explanatory goals.

Acknowledgments:

I wish to thank audiences at the 8th Quadrennial Fellows Conference (Lund, Sweden) and Exploring the Undermind conference (University of Edinburgh) in July 2016, and the British Society for Philosophy of Science Annual Meeting and Aristotelean Society/Mind Joint Sessions (Edinburgh) in July 2017 for comments on earlier versions of this paper, and two anonymous referees for this journal of the revised paper.

References

Alon, U. (2007). Network Motifs: theory and experimental approaches. *Nature Reviews Genetics* 8: 450–461.

Baronchelli, A., R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. Christiansen (2013). Networks in cognitive science. *Trends in Cognitive Sciences* 17 (7): 348–360.

Bechtel, W. (2008). *Mental Mechanisms*. New York and Oxon: L. Erlbaum.

Bechtel, W. and A. Abrahamsen (2005). Explanation: a mechanist alternative. *Studies in the History and Philosophy of Biology and Biomedical Sciences* 36: 426–441.

Bennett, M. and P. Hacker (2003). *Philosophical Foundations of Neuroscience*. Malden, MA and Oxford: Blackwell.

Bennett, M., D. Dennett, P. Hacker, and J. Searle (2007). *Neuroscience & Philosophy: Brain, Mind, and Language.* New York: Columbia University Press.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon.

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese* 191: 127–153.

Craver, C. (2006). When Mechanistic Models Explain. *Synthese* 153: 355–76.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: OUP.

Craver, C. and J. Tabery (2017). Mechanisms in Science. *The Stanford Encyclopedia of Philosophy* (Spring 2017 edition), E. Zalta, ed., URL = https://plato.stanford.edu/entries/science-mechanisms/

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge: MIT.

Dasgupta, S., C. Howcroft Ferreira, G. Meisenböck (2014). FoxP influences the speed and accuracy of a perceptual decision in Drosophila. *Science* 344 (6186): 901-04.

Dennett, D. (1975). Why the Law of Effect Will Not Go Away. *Journal of the Theory of Social Behavior* 5: 169–187.

Figdor, C. (2018). *Pieces of Mind: The proper domain of psychological predicates.* Oxford and New York: Oxford University Press.

Fodor, J. (1968). *Psychological Explanation*. New York: Random House.

Froese, T., C. Gershenson, and L. Manzanilla (2014). Can Government Be Self-Organized? A mathematical model of the collective social organization of ancient Teotihuacan, Central Mexico. *PLoS One* 9 (10): e109966.

Irvine, E. (2016). Model-Based Theorizing in Cognitive Neuroscience. *British Journal for the Philosophy of Science* 67: 143–168.

Kaplan, D. and C. Craver (2011). The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philosophy of Science* 78 (4): 601–627.

La Mettrie, J. (1748). *Man—Machine*. Trans. Jonathan Bennett, 2017. Downloaded from http://www.earlymoderntexts.com/assets/pdfs/lamettrie1748.pdf.

Leibniz, G. (1714). *The Principles of Philosophy, or the Monadology.* Sec. 17. Trans. Ariew and Garber (1989) *Leibniz: Philosophical Essays* (Indianapolis and Cambridge: Hackett): 215.

Lycan, W. (1991). Homuncular Functionalism Meet PDP. In Ramsey, Stich and Rumelhart, eds., *Philosophy and Connectionist Theory.* L. Erlbaum: 259–86.

Maienschein, J. (2017). Epigenesis and Preformationism. *The Stanford Encyclopedia of Philosophy* (Spring 2017 edition), E. Zalta, ed., URL = https://plato.stanford.edu/entries/epigenesis/

Machamer, P., L. Darden, and C. Craver (2000). Thinking About Mechanisms. *Philosophy of Science* 67 (1): 1–25.

Margolis, J. (1980). The Trouble With Homuncular Theories. *Philosophy of Science* 47 (2): 244–59.

Piccinini, G. and C. Craver (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183: 283–311.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10 (2): 59–63.

Roskies, A. (2014). Mindreading and privacy. In M. Gazzaniga, ed., *The New Cognitive Neurosciences* (Cambridge, MA: MIT Press): 1003–11.

Sober, E. (1982). Why must homunculi be so stupid? *Mind* 91: 420–422.

Stinson, C. (2016). Mechanisms in psychology: ripping nature at its seams. *Synthese* 193: 1585–1614.

Suri, R. and W. Schultz (2001). Temporal Difference Model Reproduces Anticipatory Neural Activity. *Neural Computation* 13: 841–82.

Sutton, R. and A. Barto (1981). Toward a Modern Theory of Adaptive Networks: Expectation and Prediction. *Psychological Review* 88 (2): 135–170.

Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese* 183 (3): 313–338.

Wimsatt, W. (2006). Reductionism and Its Heuristics: making methodological reductionism honest. *Synthese* 151: 445–75.