JOURNAL PRE-PROOF

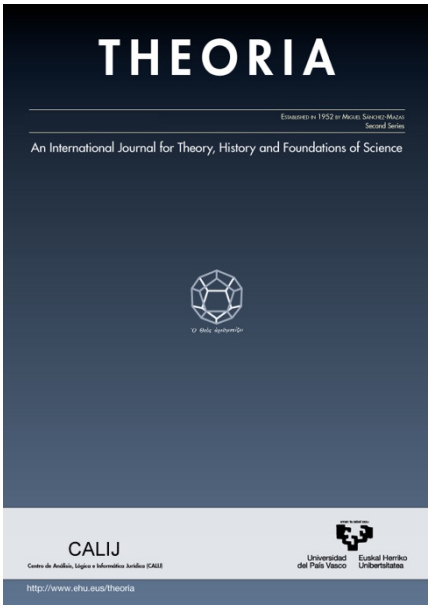
How general are Marr's levels of analysis? An assessment of Bermúdez's view

Marko Jurjako

DOI: 10.1387/theoria.25896

Received: 11/01/2024

Final version: 04/02/2025



How general are Marr's levels of analysis? An assessment of Bermúdez's view

(¿Cómo de generales son los niveles de análisis de Marr? Una evaluación de las propuestas de Bermúdez)

Marko Jurjako
University of Rijeka

ABSTRACT: In his seminal book, *Philosophy of psychology: A contemporary introduction*, José L. Bermúdez argues that David Marr's levels of analysis cannot be used as a general framework for explaining phenomena in cognitive science. More specifically, he argues that Marr's levels of analysis paradigmatically apply to subpersonal modular processes, while the mind as a whole is also characterized by non-modular cognitive systems. In this paper, I evaluate Bermúdez's arguments for this conclusion. Part of the evaluation will be based on recent advancements in the active inference framework, suggesting that the mind as a whole can be analyzed within the Marrian paradigm. Moreover, I provide some reasons for thinking that Marr's levels of analysis could also be employed to illuminate the relationship between personal and subpersonal explanations.

KEYWORDS: active inference; free-energy principle; José Bermúdez; interface problem; personal/subpersonal explanations; Marr's levels of analysis.

RESUMEN: *En su obra seminal,* Philosophy of psychology: A contemporary introduction*, José L. Bermúdez argumenta que los niveles de análisis de David Marr no pueden ser usados como un marco general para explicar fenómenos en ciencias cognitivas. Más específicamente, argumenta que los niveles de análisis de Marr se aplican paradigmáticamente a procesos modulares subpersonales, mientras que la mente como un todo también está caracterizada por sistemas cognitivos no modulares. En este artículo, evalúo los argumentos de Bermúdez a favor de esta conclusión. Parte de esta evaluación se basará en avances recientes en el marco de inferencias activas, sugiriendo que la mente como un todo puede ser analizada dentro del paradigma de Marr. Además, proporcionaré razones para pensar que los niveles de análisis de Marr también pueden emplearse para iluminar la relación entre explicaciones personales y subpersonales.*

PALABRAS CLAVE: *inferencia activa; principio de energía libre; José Bermúdez; problema de la interfaz; explicaciones personales/subpersonales; niveles de análisis de Marr.*

SHORT SUMMARY**:** This paper evaluates José L. Bermúdez's claim that Marr's levels of analysis apply only to subpersonal modular processes, not to the mind as a whole. It argues that recent advancements, particularly the active inference framework based on the free energy principle, challenge this view and demonstrate that Marr's framework can also illuminate the relationship between personal and subpersonal explanations in cognitive science.

1. Introduction[1]

Cognitive phenomena can be studied at different levels of analysis (see, e.g. Colombo & Knauff, 2020; Piccinini, 2020; Verdejo & Quesada, 2011). In this regard, David Marr (1982) provided an especially influential model for understanding cognitive phenomena (see, e.g. Peebles & Cooper, 2015; Poggio, 2012). He introduced the distinction between the computational, algorithmic, and implementational levels of analysis. Given its influence in cognitive science, many have critically discussed different aspects of Marr's levels of analysis and their relative importance (see, e.g., the papers in Peebles & Cooper, 2015; Zednik, 2018). Yet, few have explicitly discussed the capability of Marr's levels of analysis to provide a general framework for understanding psychological phenomena, including also personal and subpersonal perspectives on cognitive systems.

In a rare discussion of this issue, José Luiz Bermúdez (2005, p. 27) argues that Marr's distinction cannot be used for this purpose. He offers two related arguments to support this claim. First, he contends that Marr's distinction is best suited for explaining modular cognitive systems (see, also Cooper & Peebles, 2015). Marr's framework requires a clearly defined task and well-specified functions to generate hypotheses about a system's representations and algorithms. However, non-modular systems often lack such clear functional specifications, making them difficult to analyze using Marr's tripartite approach.

Second, Bermúdez argues that even if one successfully applies Marr's framework to a particular personal or non-modular cognitive capacity, it is very difficult—or even impossible—to extend this analysis to the mind as a whole. By this, he suggests that Marr's levels cannot provide a unified explanatory framework for understanding the general relationship between different levels of explanation in cognitive science, particularly between personal and subpersonal perspectives on the mind/brain.

The aim of this paper is to evaluate Bermúdez's claim that Marr's levels of analysis cannot provide a general framework for explaining cognitive processes. To address the first argument, I will show that cognitive (neuro)science offers computational models of higher-order, non-modular systems that can be usefully analyzed using Marr's tripartite distinction. To address the second argument, I will argue that Marr's framework can be applied to the mind as a whole and even shed light on the relationship between personal and subpersonal explanations, provided that a sufficiently rich theoretical framework is adopted. As a proof of concept, I will show that the active inference framework, grounded in the free-energy principle, provides a unified framework for understanding the overall function of the mind and its physical implementations. Given its computational structure, this framework lends itself well to analysis through Marr's levels.

The paper proceeds as follows. Section 2 summarizes Marr's levels of analysis. Section 3 contextualizes the debate on the generality of Marr's framework within Bermúdez's "interface problem". There I assess Bermúdez's first argument that Marr's levels of analysis are limited because they mostly apply to subpersonal and modular cognitive processes. Section 4, in a preliminary way,

[1] This paper builds upon and provides a more comprehensive elaboration of some of the ideas initially presented in Jurjako (2023).

examines Bermúdez's second argument, which claims that Marr's levels of analysis, even if applicable to some higher-level non-modular processes, are inadequate for addressing the function of the mind as a whole. In Section 5, I defuse this argument by relying on the active inference framework, arguing that it provides sufficient conceptual resources for understanding the function of the mind as a whole. Section 6 revisits the interface problem, exploring how the application of Marrian analysis within the active inference framework can clarify the relationship between personal and subpersonal explanations.

2. Marr's levels of analysis

Marr (1982) famously argued that a complete understanding of some cognitive capacity involves an analysis at computational, algorithmic, and implementational levels. To illustrate how this framework is used, let us consider the human ability to visually recognize objects. At the computational level, we are trying to determine what the system under consideration is doing and why, i.e., what is the function it is supposed to perform.[2] We can say that the function of visual object recognition is to recognize objects by identifying and classifying them based on their visual features, such as shape, size, color, and texture. In computational terms, this function of visual object recognition can be framed as a process in which the goal is to map the visual input (e.g., a 2D image on the retina) to a representation of a 3D object that is classified under some label (e.g., a cat or a dog).

At the algorithmic level, the aim is to determine how the system implements the functions and tasks identified at the computational level. This involves determining an algorithm that describes the steps and procedures that the system uses to solve the identified computational problem, and to identify the representations it uses to achieve this goal.[3] In the case of visual object recognition, an algorithm for identifying and classifying 3D objects could use procedures for filtering visual inputs to detect edges of objects. As edges of objects are correlated with sudden changes in the intensity of the light that is reflected by an object, they can be represented by mathematical procedures that use zero-crossings. Zero-crossings are the points in an image where the intensity of the pixels changes from positive to negative values, or *vice versa*. Then, the algorithm may use procedures that extract features such as shape, size, and texture of an object from the zero-crossings. Finally, to classify the object, the algorithm may compare the extracted features with information about objects that is already stored in memory.

The implementational level aims to specify the neural structures and mechanisms that underpin algorithms involved in visual object recognition. For example, it has been proposed that the visual system uses a hierarchy of processing stages, with early stages dedicated to basic features such as edges, and later stages dedicated to more complex features such as shape, size, and texture.

[2] For a seminal discussion of Marr's computational level, see Shagrir (2010); for a recent application, see Mallory (2024).

[3] Marr is working in the classical paradigm of cognitive science that understands the mind based on the digital computer metaphor. Thus, it was natural for Marr to talk about representations and algorithms for transforming representations in a way that computing devices, such as cashier registers, manipulate representations. However, nothing in the idea of the algorithmic level as conceived by Marr commits one to adopt the idea of digital representations that are manipulated according to their structural features (see, e.g. Zednik, 2018). The minimal presupposition for the application of the Marrian paradigm is that cognitive processes might be interpreted as information processing devices, without specific commitments about how a cognitive system under examination represents information and what types of equations are used to describe their transformations.

Some of the specific neural structures involved in this process are the retina, the primary visual cortex (V1), and the inferotemporal cortex (IT) (see, e.g. DiCarlo et al., 2012).

Marr's levels of analysis were devised for analyzing cognition understood as an information processing system (Poggio, 2012). It is unclear whether this analysis can be used for understanding the general relation between different levels at which psychological phenomena can be explained, including the relation between the personal and the subpersonal. This question has been surprisingly underexplored in the literature. In a rare discussion of this issue, Bermúdez contends that Marr's levels of analysis are unsuitable for this task, as will be discussed in the next section.

3. Can Marr's distinction be used as a general framework for analyzing phenomena in cognitive science?

3.1. PERSONAL AND SUBPERSONAL EXPLANATIONS, AND THE INTERFACE PROBLEM

Bermúdez situates his arguments within the discussion of the relation between personal and subpersonal explanations (see Bermúdez, 2005, pp. 17-38). There is some uncertainty about how to clearly distinguish between the personal and the subpersonal (for recent discussions, see Dänzer, 2023; Drayson, 2012; Westfall, 2022). Nonetheless, for the purposes of our discussion, Bermúdez offers a reasonably clear approach to conceptualizing the distinction.

In line with a venerable tradition in philosophical psychology, Bermúdez understands personal explanations as referring to traits that persons, or more broadly, rational agents, can properly possess (Dennett, 1969; Hornsby, 2000). Specifically, he claims that the personal is characterized by three features (see Bermúdez, 2000, p. 64). The first feature is its distinctive vocabulary, which refers to intentional states such as beliefs, desires, and intentions. The second feature is that personal explanations individuate a special class of regularities that might not be discernible from other perspectives. For instance, personal explanations often reference intentional actions, such as intending to make a money transfer, which can be realized in various physical ways—like making hand gestures, giving a verbal command, or other movements that share nothing in common except being instances of the same regularity recognized as "transferring the money" (see, also Fodor, 1974). The third feature of the personal level is that its explanations are guided by principles of rationality (Davidson, 2001; Dennett, 1981). For instance, in explaining why someone transferred money, we might assume that they believed the transfer would settle a debt and wanted to resolve it, thus choosing a method—like signing a check or making an electronic payment—that would achieve that goal (for influential discussions of rationality and its various applications across the social sciences and humanities, see Bermúdez, 2011, 2020).

Bermúdez also suggests that these features characterize the folk-psychological explanations we use in everyday life for social interaction.[4] We typically understand people as possessing intentional states and assume that they act to satisfy their goals in light of their beliefs, in a minimally

[4] Recent scholarship suggests that associating personal and folk-psychological explanations only with intentional states and rational principles is overly simplistic. Such explanations also include personality traits, habitual actions, social scripts, and references to skills that do not necessarily involve intentional states or rational constraints (Andrews et al., 2021; Westfall, 2022; see, also Bermúdez, 2005, ch. 7). However, for the purposes of this paper and formulating the interface problem, Bermúdez's three-feature characterization of the personal and its relation to folk psychology will suffice.

rational way. He further notes, however, that the personal level cannot be entirely equated with folk-psychological explanations, as the personal perspective is also employed in cognitive psychology. In this field, scientists measure reaction times, use verbal reports, and formulate hypotheses about conscious cognitive and perceptual abilities, thereby contributing to the development of scientific theories of personal-level states and processes (Bermúdez, 2000, p. 65). While it is reasonable not to equate personal explanations entirely with folk-psychological ones, we will set this distinction aside in the following discussion. This is because scientific psychology's personal-level constructs can be seen as more precise extensions of folk-psychological constructs and associated principles (see Bermúdez, 2005, p. 65). Moreover, insofar as the personal level is characterized by folk-psychological explanations and constructs, it becomes easier to frame the interface problem as Bermúdez does. However, a proper understanding of the interface problem requires a clear contrast between personal and subpersonal explanations.

The subpersonal is typically associated with states, activities, and processes that operate below the level of a person. Although subpersonal constructs are in some sense related to those we apply at the personal level, they should be clearly distinguished (see Bermúdez, 2005, ch. 2; Drayson, 2012). For example, Marr's theory of vision is commonly regarded as a subpersonal theory of visual processing. As explained in Section 2, according to Marr, the ultimate task of early vision is to produce a 3D representation of the external scene. However, the concept of representation in Marr's account should be clearly distinguished from the notion of representation of a visual scene we attribute to whole persons. The main difference is that unlike personal representations, which are typically consciously entertained, Marr's representational processes operate subconsciously.[5] Moreover, Marr's theory of vision provides computational principles intended to explain how subconscious representations in early vision are processed to eventually yield a representation of the external environment that our cognitive system might further use for object recognition and classification. At the personal level, however, we are concerned with reasoning processes that rely on conscious perceptions, which contribute to forming propositional attitudes and provide inputs to our decision-making processes characteristic of intentional agency.

Subpersonal explanations are also linked to important distinctions in how different disciplines approach the study of the mind (Jurjako, 2024). The personal perspective is used when we consider mental states, processes, and abilities attributed to people understood as agents capable of rational thought, while the subpersonal perspective is typically applied in cognitive science and neuroscience, where people's perceptual and cognitive abilities are understood as products of mechanisms, where they can be explained at different levels of analysis, involving their functions, causal roles, and structural composition (Shea, 2018). In this respect, the subpersonal level can be characterized by various levels of analysis typically studied in cognitive (neuro)science. These include

[5] It is worth noting that sometimes, the primary criterion for individuating the personal level involves references to conscious states and processes, while the subpersonal level is understood as pertaining to subconscious states and processes. However, although the personal perspective often refers to conscious states—and rightly so, since our conscious experiences permeate personal-level functioning (Chappell, 2023)—its explanations and associated states cannot be individuated solely by the criterion of consciousness. This is because there may be subconscious beliefs attributed to a person, and we may also explain people's behavior by referring, for instance, to character traits that do not necessarily operate at the conscious level (Westfall, 2022; see also the discussion in Bermúdez, 2005, pp. 29-31).

cognitive-computational processes as explored in cognitive psychology, analyses of how populations of neurons enable cognitive and motor capacities, detailed explanations of individual neuronal responses to specific stimuli, and even more fine-grained aspects studied in molecular neuroscience. Each of these levels contributes to our understanding of how subpersonal processes underlie complex cognitive and behavioral phenomena (see Bermúdez, 2022).

According to Bermúdez, considering the personal and various levels of subpersonal explanation leads to the question of how to understand their relationship. This is what he calls the interface problem:

> This is the problem of explaining the relation between the commonsense, everyday type of psychological explanation that we all engage in every day […] and the levels of explanation lower down in the hierarchy. (Bermúdez, 2005, p. 35)

To further illustrate the interface problem, he articulates it through a series of questions:

> How do explanations of the behavior of people given in terms of their beliefs, desires and other psychological states mesh, for example, with explanations in terms of patterns of activity across populations of neurons? How does the biochemistry of what goes on inside a neuron relate to the dynamics of how a person interacts with the environment? What is the relation between understanding a person as a conscious, reasoning agent, on the one hand, and understanding that person's brain as a complicated type of computational mechanism? (Bermúdez, 2005, p. 35)

As we can see here, Bermúdez formulates the interface problem as one of understanding a hierarchy of explanations, where personal explanations occupy the upper levels of the hierarchy and subpersonal explanations capture the lower levels. According to his view, one of the main tasks of the philosophy of psychology is to provide a coherent model for understanding the relationship between these levels (Bermúdez, 2005, p. 35). Given that Marr's levels of analysis offer a clear approach to understanding and explaining cognitive phenomena, it is worth considering whether this framework can sufficiently illuminate the relationship between these levels of explanation.

At first glance, it seems that the interface problem could be straightforwardly addressed from a Marrian perspective.[6] Marr's levels of analysis—delineated by the questions "why", "how", and "where"—can be seen as aligning with the levels of explanation Bermúdez discusses. In this view, the personal level, which largely involves folk-psychological explanations, could be understood as providing explananda with functions specified at the computational level. The algorithmic level would then seek to identify the algorithms and representations that give rise to these psychological functions. Finally, the implementational level would investigate their biological underpinnings.

[6] It is worth noting that an often-overlooked question is whether the interface problem should be addressed uniformly across cognitive science or if different fields require distinct solutions. I touch on this issue in a recent philosophy of psychiatry paper, where I discuss the notion of mental disorder in relation to the interface problem (see Jurjako, 2024).

However, Bermúdez offers reasons for doubting that this straightforward solution can work. In the following sections, we will examine the cogency of his skepticism regarding the use of Marr's levels of analysis for addressing the interface problem.[7]

3.2. MODULAR AND NON-MODULAR SYSTEMS, AND MARR'S LEVELS OF ANALYSIS

One reason for thinking that Marr's levels of analysis are particularly limited in explaining cognitive phenomena is that they only pertain to modular systems. In this regard, Bermúdez writes that "it should be clear that nothing like Marr's account could be *straightforwardly* applied to what we might think of as higher (i.e. non-modular) cognitive processes" (Bermúdez, 2005, p. 27). The reasoning seems to be as follows. Marr's levels of analysis are well suited for explaining modular systems. Here Bermúdez (2005, p. 25) employs Jerry Fodor's classical (1983) notion of modularity.[8] According to this view, modular systems are, among other things, domain specific (they are specialized for well-specified tasks) and informationally encapsulated (they only respond to a finite set of inputs and cannot be penetrated with other types of information). This implies that modular systems possess clear functional specialization with evident implementability through well-defined algorithms. In contrast, personal states are typically associated with non-modular processes. This apparently means that they will lack clearly specified functions that can be algorithmically implemented. Thus, typically it would not be feasible to apply Marr's analysis to personal non-modular processes.

Even though Marr's levels of analysis well capture modular processes, it does not follow that they are not useful for explaining non-modular processes. Consider, for instance, our ability for ordinary reasoning. This ability is informationally "promiscuous", in the sense that it can take in information from different cognitive systems, integrate it, and make further inferences based on it. Moreover, such an ability is not domain-specific, at least not in the sense that it is restricted to responding to a narrow set of inputs. For example, a person can by seeing the weather outside form the belief that it is sunny outside and by hearing the rain form the belief that it is raining outside. By integrating these beliefs, the person can conclude that it is sunny and raining at the same time.

More importantly, such a non-modular ability can be usefully analyzed at the computational, algorithmic, and implementational levels. For example, at the computational level, this ability can be understood as the ability for reasoning that takes symbolic representations as input and outputs other symbolic representations. At the algorithmic level, we can think about different representational systems and associated rules for their processing. For instance, we can formulate algorithmic procedures based on Bayesian statistical inferential procedures or inferential procedures based on some form of defeasible logic. Finally, we can consider where and how these logical

[7] In what follows, I will refer to Bermudez's argument as presented in his (2005) book. The same arguments are also discussed in his cognitive science textbook (see 2014, pp. 126-129). Interestingly, in the last (fourth) edition of the textbook, the discussion of these arguments has been largely omitted.

[8] For the sake of discussion, I follow Bermúdez in assuming the Fodorian view of modularity. There are other non-Fodorian perspectives on how modular processes could be understood. In the next section, I will introduce the active inference framework, which is often understood as challenging the sharp distinction between modular and non-modular systems (for discussion, see Drayson, 2017). Given that this issue is not directly pertinent for the present discussion, it will not be further examined.

procedures are implemented in the brain (see, e.g. Baggio et al., 2015). Thus, apparently non-modular systems can have well-specified functions that are amenable to Marrian analysis.

Against such a piecemeal response, Bermúdez offers what I take to be his second argument for the limitations of Marrian analysis. This argument takes the form of a more general objection.[9] He asserts that

> [E]ven if [Marr's levels of analysis] could be extended to non-modular processes, […] it will certainly be *impossible* to do so for the mind as a whole—and it is, of course, an understanding of the mind as a whole that we are ultimately aiming for. Marr's analysis of the early visual system provides a clear illustration of the general idea of a hierarchy of different levels of explanation. But it is not itself pitched at the right sort of level to provide a model of how we might understand the general idea of a hierarchy of explanation applied to the mind as a whole. (Bermúdez, 2005, p. 27, emphasis added)

There appear to be two distinct issues in Bermúdez's quote. The first relates to the claim that Marr's levels of analysis cannot be applied to the mind as a whole. We will explore Bermúdez's reasoning for this in the next section. The second issue pertains to Bermúdez's assertion that Marr's levels are unsuitable as a model for understanding "the general idea of a hierarchy of explanation applied to the mind as a whole". These two claims seem connected as a premise (first claim) leading to a conclusion (second claim). Specifically, if we ask why Bermúdez claims that Marr's levels of analysis do not provide a suitable framework for explaining the mind as a whole, the answer seems to be that this is because the mind, as an integrated system, cannot be fully analyzed through Marr's levels.

The more interesting claim seems to be the first one, and later on I will focus on it. This is because even if Bermúdez is correct that Marr's levels of analysis do not offer a clear model of the hierarchy of explanations in cognitive science, this would not be particularly surprising or controversial. Indeed, there is general agreement among philosophers of cognitive science that Marr primarily proposed levels of *analysis* of cognitive phenomena, without specifying how these levels should align with different types of explanations applied to cognitive phenomena. To address this gap, several philosophers have suggested connecting Marr's levels of analysis with mechanistic explanations (see, e.g. Bechtel & Shagrir, 2015; Zednik, 2018). According to such approaches, Marr's levels of analysis can naturally be understood as applying to mechanisms underlying cognitive phenomena of interest by guiding us to ask questions about a mechanism's function, how it performs that function, and where it is physically implemented. Given that mechanisms are

[9] In addition to the following objection, Bermúdez (2005, pp. 26-27) also suggests that Marr's levels of analysis might be limited to modular processes because algorithmically analyzing non-modular processes might face the frame problem—a challenge in cognitive science and AI where a system struggles to determine which aspects of an ever-changing environment are relevant to its current goal, making it difficult to algorithmically handle complex, context-sensitive tasks. However, this objection does not seem too serious. While Bermúdez is right that the frame problem poses a challenge for creating algorithms for non-modular processes, this is not a principled problem for applying the Marrian analysis because its usefulness does not depend on whether we can at this moment invent algorithmic procedures that would be able to model every aspect of the mind. Theorizing and practical work in cognitive science starts with the presupposition that the human mind has somehow solved the frame problem. And by studying the mind from the computational, algorithmic, and implementation levels, we are trying to discover how our minds are capable of non-modular higher-level thinking and where these processes are implemented in physical systems such as the brain.

composed of parts and processes that can be structured at different levels, interpreting Marr's proposal within a mechanistic philosophy of science offers a clear view of how to formulate a Marrian-inspired model of explanatory hierarchy within cognitive science (for further discussion, see Chirimuuta, 2024; Zednik, 2018).

However, if Bermúdez is correct that Marr's levels of analysis cannot be applied to the mind as a whole, then it would follow that no matter how detailed our Marrian mechanistic explanations are, they would be lacking in certain respect, and would not enable us to provide a coherent framework for thinking about cognitive phenomena spanning different levels of description. For this reason, in what follows, I will focus on evaluating Bermúdez's claim that Marr's levels of analysis cannot be extended to characterizations of the mind as a whole.[10]

4. Why Marr's levels of analysis cannot be applied to the mind as a whole?

It is not immediately clear why Bermúdez thinks that a Marrian analysis cannot be applied to the mind as a whole. One reason that emerges in Bermúdez's discussion relates to the claim that Marr's analysis cannot be applied to people as they are understood at the personal level of functioning. This line of reasoning begins with a reasonable claim made by Bermúdez that

> Cognition is not an isolated activity and if we are interested in studying the mind as a whole we must start from the twin facts, first, that it is organisms that have minds and, second, that possessing a mind allows those organisms to behave in the ways characteristic of intelligent agents. (Bermúdez, 2005, p. 28)

The idea seems to be that to explain and understand the role of the mind as a whole, we must pay heed to its role in regulating the behavior of cognitive agents. People are paradigmatic cognitive agents. Thus, the explanation of the mind must begin with an explanation of the behavior of the person. However, according to Bermúdez:

> Theories such as Marr's operate at a lower level than the level of cognitive agents. They deal with parts or modules of the cognitive agent, rather than with the agent itself as a thinking and acting organism. They are theories at the subpersonal level (below the level of the person). (Bermúdez, 2005, p. 28)

While it is true that Marr's specific theory of early visual processing is best understood as a subpersonal theory, this does not imply that his framework for levels of analysis cannot be extended to other, non-subpersonal cognitive phenomena. Indeed, as argued earlier, it seems we can apply Marr's levels of analysis to a paradigmatically non-modular personal-level ability, such as the ability for making ordinary/logical inferences? Why, then, should we assume that Marr's levels of analysis are limited in this regard?[11]

[10] Thanks to an anonymous reviewer for pressing me to more adequately delineate the scope of this discussion.

[11] Richard Cooper and David Peebles (2015, p. 253) also notice that Marr was primarily applying levels of analysis to components and subcomponents of cognitive abilities, and not to whole agents. However, despite this observation, they do not claim that Marr's levels cannot in principle be applied to whole agents.

There seems to be something about the concept of the mind as a whole that Bermúdez believes resists analysis in Marrian terms. The claim seems to be that if the mind as a whole cannot be provided a functional specification, then attempting a computational analysis would be futile and we would not be able to formulate explanations of how the mind actually performs these tasks or where these processes are implemented. Indeed, in this regard Bermúdez writes:

> [W]hen we are thinking about the mind as a whole, there are difficulties applying the type of functional analysis that Marr applied to the early visual system. When we are thinking about the mind as a whole it is very difficult, and perhaps even impossible, to identify tasks that can be understood in a determinate enough way to yield algorithms. (Bermúdez, 2005, p. 28)

Essentially, Bermúdez seems to argue that using Marr's levels of analysis to provide a general framework for explaining cognitive phenomena spanning across personal and subpersonal levels is virtually impossible because the mind, as an integrated system, lacks a clear functional specification.

It is not entirely clear why Bermúdez believed that the mind as a whole lacks functional specification. Perhaps, at the time of his writing, philosophers of cognitive science struggled to conceptualize how non-modular cognitive systems, with their domain generality and cognitive penetrability, could be computationally analyzed.[12] Be that as it may, I will not focus on exploring additional reasons for this view. Instead, I will shift my approach and offer a proof of concept. Specifically, I will argue that there are no fundamental obstacles to using Marr's levels of analysis to understand the mind as a set of abilities characterizing the behavior of whole organisms—at least, not if one adopts a sufficiently ambitious framework for conceptualizing the mind. Additionally, I will argue that working within a sufficiently rich cognitive paradigm can shed light on the relationship between personal and subpersonal levels of functioning. To support this, I will introduce the active inference framework in the next section and illustrate how Marrian analysis can be applied within this paradigm to enhance our understanding of the mind as a whole.

5. Active inference and the mind as a whole

5.1. THE ACTIVE INFERENCE FRAMEWORK[13]

The active inference framework is based on the Free Energy Principle (FEP). According to the FEP, the general function of the mind can be understood as minimizing free energy (for an overview, see, e.g. Mann et al., 2022).[14] More generally, the FEP prescribes the general conditions for self-

[12] In this regard, Bermúdez might have been influenced by Fodor's (1983) argument that there could never be a cognitive science of non-modular systems. Such systems are characterized by domain generality and holism, where any piece of information can influence any other cognitive process, leading to intractability (for discussion, see Murphy, 2019, see also footnote 8 above).

[13] In what follows, I provide a basic introduction to the formal machinery of active inference and the Free Energy Principle. An informed reader familiar with this literature may skip ahead to Subsection 5.2.

[14] Some authors differentiate between high and low road approaches when introducing the active inference framework (see Parr et al., 2022 chs. 2-3). The low road approach involves tracing the development of active inference from Helmholtz's idea of perception as inference all the way to contemporary predictive processing accounts, according to which the brain promotes adaptive behavior by minimization of prediction errors (for a notable book-length discussion, see Hohwy, 2013). In contrast, the high road approach involves the normative perspective, starting with first principles concerning the necessary conditions for organisms to maintain their existence, which in this case involves minimization

organizing systems to remain in existence. The survival of all self-organizing systems, such as living organisms, depends on their ability to resist the tendency to dissipate under the environmental pressures. The FEP states that self-organizing systems maintain their integrity and resist dissipation by minimizing atypical or surprising events in their environment. Granting the insight that organisms possess minds that enable intelligent behaviors, it follows in the present context that we can construe the general function of the mind as the ongoing process of minimizing uncertainty, which will later be expounded in terms of free energy minimization.[15]

To better understand the basic notions underpinning the FEP, we need to elaborate on what is meant by surprise, self-organizing systems, and the role of free-energy in minimizing surprisal. We will start with the formal notion of surprise.

In information theory, surprise is a measure of the amount of unexpectedness associated with a particular event. To distinguish it from the commonsensical notion of surprise, researchers tend to call this quantity surprisal. It is defined as the negative logarithm of the probability of an event occurring:

$$\text{Surprisal: } -\log P(s) \qquad (1)$$

Here, $P(s)$ denotes probability of *s,* –log is a logarithmic function of the probability function. Surprisal is inversely proportional to the probability of that event occurring. In other words, the higher the probability of an event, the lower the surprisal, and *vice versa*.

The general idea is that adaptive action involves remaining in familiar states by minimizing surprisal or uncertainty (see, e.g. Buckley et al., 2017). For instance, for a fish, familiar states with low surprisal include being in water, while being outside water involves being in states with high surprisal. In the basic case, the surprisal of a state is determined by how expected it is for an organism to be in that state from the perspective of its phenotype and the ecological niche it inhabits. Thus, in this context, surprisal is determined by the conditional probability that an organism will experience an event or be in some state given its phenotype. From a formal perspective, we can think about an organism's phenotype as embodying a model of its relations with the environment it is embedded in. This leads us to the idea of a Markov blanket which determines the boundaries between self-organizing systems and their surrounding environment.

According to the FEP, Markov blankets determine the identity of a self-organizing system. A Markov blanket is a statistical construct originally introduced to capture efficient and qualitative forms of probabilistic inference (Pearl, 1988). The goal was to capture relevance and dependency relations among variables via probabilistic graphical models. In general, this is accomplished by

---

of free energy. For the present purposes, introducing the framework from the high-road perspective seems preferable as it will enable us to more perspicuously think about the function of the mind as a whole.

[15] An anonymous reviewer for this journal has rightly suggested that the claim in this section—that the framework based on the FEP provides a coherent way to conceptualize the mind as a whole—can be connected to the work of Julian Kiverstein and Matt Sims (2021). They argue that FEP can offer a "mark of the cognitive", providing criteria to distinguish organisms whose behavior can be explained in cognitive, rather than non-cognitive, terms. While this connection is interesting and warrants further exploration, it lies beyond the scope of this paper. Our focus here is on whether Marrian analysis can be applied to the mind as a whole, rather than on distinguishing organisms by attributing cognitive processes.

identifying the set of variables that constitute the Markov blanket of the target variable. The Markov blanket includes the parents (direct causes) of the target variable, the children (direct effects) of the target variable, and any other variables that are probabilistically connected to the target variable, while excluding the conditionally independent variables (i.e. those that do not enable inferring the target variable), and thus ensuring that all relevant information and dependencies are captured.

In the context of FEP, Markov blankets are used to formalize the partitioning of a self-organizing system into its internal, external (environmental), and the boundary or blanket states that separate them (see, e.g. Hipólito et al., 2021). The idea is that the internal states are separated, i.e. conditionally independent from the external states given the boundary states. The boundary states are further divided into sensory and active states. The relation of dependence is such that external states can influence sensory states that further cause changes in internal states. In contrast, the internal states can cause changes in active states that in turn cause changes in the external states.[16]

These relations form the core of the *active inference* framework. Sensory states represent perceptual processes that affect an organism's internal states, while active states play the role of actions that unidirectionally stem from internal states and cause changes in the external state. Together, the functioning of sensory and active states form causal perception-action loops that enable the organism and its internal states to efficiently adapt to the continuously changing environment and in this way to minimize surprisal. To adapt to the external environment an organism needs to be able to detect the external states and accordingly adjust its internal states or adjust the external states to make them conducive to its well-being. Adjusting internal states is called perceptual inference, while adjusting external states via action is called active inference.

Given the existence of a Markov blanket, surprisal can be minimized only indirectly, via the influence of sensory and active states because only through them can the internal and external states exchange energy, matter, and information (Hipólito et al., 2021). In probabilistic terms, this means that the organism faces the problem of determining the posterior probability of external states given its blanket states. Formally, problems of this type are solved by employing Bayesian inference. To implement computational procedures that produce results approximating Bayesian inference, it has to be supposed that the brain (implicitly) embodies a generative model of how the external states causally affect its sensory states.[17]

[16] It should be noted that biological systems at different levels of spatiotemporal organization can be construed as having a Markov blanket (Hipólito et al., 2021). For instance, the membrane of a cell can be conceived as a boundary of the cell's Markov blanket, without which the cell would dissipate and fall out of existence. Similarly, in the brain neurons can be understood as entities with Markov blankets, but also populations of neurons, whole brains, and entire organisms. Given this flexibility about which entities can be construed as having Markov blankets, and therefore fall under the purview of active inference, some have criticized the whole framework as not being sufficiently restrictive about delineating cognitive systems (for discussion, see Bruineberg et al., 2022). I will not delve into the ontological disputes of the active inference framework. For my purposes, it is sufficient to think of Markov blankets as characterizing entire organisms and the neurobiological processes that underpin the relevant mental capacities.

[17] It should be added that the debate is still ongoing whether the discourse about generative models and probabilistic Bayesian beliefs should be understood in realist or instrumentalist terms (for discussion, see Kirchhoff et al., 2022; Kiverstein & Sims, 2021; Ramstead et al., 2020). As this does not directly affect the discussion in the paper, I will remain neutral about this issue.

The role of generative models is to represent the generative processes, that is, the causal structure of the external states and how they affect sensory states. In formal terms, a generative model is a joint probability distribution of two variables:

$$P(h, s) = P(s \mid h) * P(h) \qquad (2)$$

Here, h is a variable of the internal model that represents a hypothesis about how sensory data s are generated by external processes. The generative model can be decomposed into the prior probability of h and the likelihood that s would be observed under the hypothesis h. Prior probabilities and likelihoods are often called beliefs that the brain has about the external generative processes.[18] Given this generative model, the activities of the brain can be cast as implicitly engaging in inferential processes that compute the posterior probability of the hypothesis via Bayes theorem:

$$P(h \mid s) = (P(s \mid h) * P(h)) \,/\, P(s) \qquad (3)$$

Importantly for our context, it should be noted that performing Bayesian inference involves maximizing model fit with the available evidence or data. Moreover, maximizing model fit (i.e. $P(s)$) is equivalent to minimizing surprisal of the model (i.e. $-\log P(s)$). From this follows that Bayesian inference is equivalent to information processing that minimizes surprisal of s (see Parr et al., 2022, pp. 18-19).

The surprisal of the brain's sensory states can be minimized by Bayesian inference (or alternatively, its model evidence can be maximized) in two ways. The first is by performing perceptual inference where surprisal is minimized by sensory states. For instance, a person's generative model may ground an expectation (i.e. a prior belief) that a glass of water is on the table, but the incoming sensory data do not confirm this expectation. To reduce this surprisal, the person can update their generative model that the glass is not on the table. Alternatively, active states can reduce surprisal. For instance, the person could reduce surprisal by putting the glass of water on the table and thus changing the world instead of their sensory states.

However, estimating surprisal and minimizing it by Bayesian inference is often computationally intractable (see Parr et al., 2022, p. 27). Surprisal depends on marginal probabilities of evidence and the more complex generative models get, the more variables need to be marginalized out to calculate the probability of evidence, which makes Bayesian inference intractable. This problem can be solved by using variational procedures that approximate optimal Bayesian inference. This is where free energy minimization becomes relevant.

According to FEP, the quantity that organisms actually minimize is variational free energy. In information theory, free energy provides an upper bound to an event's surprisal. Given that free energy is more feasible to estimate than surprisal, the basic claim of the active inference framework is that organisms minimize state uncertainty (i.e. their surprisal) by minimizing variational free

[18] This talk about "beliefs" should not be confused with beliefs as personal level states (see Dewhurst, 2017). Here, "beliefs" denote probability distributions that compose the generative models, and as such are typically taken to characterize the brain's subpersonal states and processes.

energy. Formally, this means that posterior beliefs are replaced by a distribution that approximates them, called approximation posterior Q, while log probability of evidence is replaced by variational free energy F. Thus, optimal Bayesian inference is formulated as minimization of variational free energy F. Free energy is a functional $F[Q, s]$ that takes as arguments approximate posterior Q and sensory signals s (that represent available evidence for the generative model). There are several equivalent mathematical formulations of the free energy minimization principle (see Parr et al., 2022, pp. 28-30). We will focus on the one that makes clear how minimizing free energy approximates minimization of surprisal:

$$F[Q(h), s] = DKL[Q(h) \,||\, P(h|s)] - \log P(s) \qquad (4)$$

The first term is the Kullback-Leibler (KL) divergence between the approximate posterior distribution Q of a hypothesis h and the true posterior probability of the hypothesis given the sensory input (i.e. $P(h|s)$). The KL divergence measures how different these two distributions are from each other. Minimizing this term means making $Q(h)$ as similar as possible to the true posterior $P(h|s)$. Minimization of KL is usually construed as involving perceptual inference. If perceptual inference reduces the KL divergence to zero, free energy would be equal to surprisal (i.e. $-\log P(s)$). The second term in the equation is the already familiar negative log probability of sensory evidence s that measures the surprisal of model evidence, i.e. how well the model's predictions match the observed data. Minimizing this term means maximizing the likelihood of the observed data given the model's predictions. This can be directly achieved only via active inference, i.e. by preforming actions that affect which sensory evidence is available to the agent. In a nutshell, given that free energy F is an upper bound to surprisal, equation (4) shows that by minimizing the divergence between the approximate posterior $Q(h)$ and the true posterior $P(h|s)$ (via perceptual or active inference), the free energy of a generative model gets constantly closer to its surprisal (i.e. $-\log P(s)$).

In summary, according to the active inference framework, action and perception present two sides of the same coin, whose ultimate function is to minimize free energy of a generative model. In more familiar philosophical terminology, perceptual inference exemplifies a mind-to-world direction of fit (i.e., the brain adjusts its internal models to match the external reality it perceives), while active inference can be associated with a world-to-mind direction of fit (i.e., by initiating action that brings the environment in line with brain's internal expectations). With these considerations in view, in the next subsection, we turn to the task of explaining how within this framework Marr's levels of analysis can be fruitfully used to analyze the mind as a whole.

5.2. MARR'S LEVELS OF ANALYSIS AND THE ACTIVE INFERENCE FRAMEWORK

5.2.1. The computational level

At the computational level, the FEP provides a general view according to which the function of the mind is minimization of free energy. Indeed, at the most general level, FEP is typically understood as a normative framework that captures "*what* living organisms must do to face their fundamental existential challenges (minimize their free energy) and *why* (to vicariously minimize the surprise of their sensory observations)" (Parr et al., 2022, p. 8, emphasis in the original text). For organisms to

survive and remain in homeostatic states they need to minimize surprisal, which is accomplished by minimizing free energy. The crucial component for achieving this is the supposition that the organisms' brains possess a "*generative model* that describes the problem the brain is trying to solve" (Parr et al., 2022, p. 105, emphasis in the original text). Once this model is identified, by performing active inference the researcher can derive predictions about the behaviors, inferences, and associated neural dynamics, and compare the model with the attained data.

By providing a comprehensive and unified perspective on the mind, the FEP also offers conceptual tools to contemplate the connections between personal and subpersonal levels. As mentioned before, the FEP casts perceptual, cognitive, and behavioral abilities as different ways of accomplishing the same thing, namely minimizing free energy. In this regard, higher-level mental states and processes (including beliefs, desires, intentions, etc.) can be construed as different aspects of ways in which organisms perceive, act, and maintain homeostasis by minimizing free energy (Friston et al., 2017). Moreover, applying the FEP to the mind as a whole suggests a deep continuity between personal and subpersonal levels. By framing higher-level mental states as manifestations of the same fundamental drive to minimize free energy, the FEP suggests that our conscious experiences, and other mental states as captured by folk-psychology, emerge from the very same principles that govern subpersonal processes (Smith, Ramstead, et al., 2022). Thus, the active inference framework, in a sense, blurs the categorical distinction between personal and subpersonal levels by portraying them as distinct means of accomplishing the same computational tasks that enable organisms to survive and maintain homeostasis.[19]

5.2.2. The algorithmic level

At the algorithmic level of analysis, this framework provides insight by offering process level theories of cognitive phenomena (Friston et al., 2017). Process theories encompass explanatory frameworks that describe more abstract mechanistic components underlying belief updating in the brain, as well as its broader impact on an organism's interactions with the environment (see Parr et al., 2022, p. viii). The development of process-level theories involves elaborating on the specifics of a generative model that is designed to capture the cognitive or behavioral ability under investigation. This is achieved by considering several factors (see Parr et al., 2022, pp. 106-113):

1. Markov blanket: firstly, we need to decide which system we are modelling. This is done by determining the Markov blanket of the system under investigation, including the determination of the internal, external, and the interfacing blanket states. For instance, the system might be the organism as a whole, but also its parts, such as the brain, brain parts, or other bodily or extra body regulatory systems that might be construed as performing inferences by minimizing free energy (see, e.g. Hipólito et al., 2021; Parr et al., 2022, p. 109).
2. Representational primitives: secondly, we need to choose the form of the generative model. This includes deciding on the appropriate type of variables, parameters, and the depth of the spatio-temporal hierarchy over which inferences and learning are defined. For instance, modelling perceptual processes that involve luminance contrasts would involve continuous

[19] In the next section, I provide further considerations on how active inference can be employed to elucidate the relationship between the personal and subpersonal levels.

variables, while modelling processes underlying object recognition would involve categorical variables.

3. Components of the generative model. Thirdly, we need to decide which parts of the model are fixed and which need to be learned. This includes determining the changeable parts of the model that are determined by the data, and priors that may be supplied by the researcher. Moreover, depending on the type of learning we are modelling, there will be a distinction between variables whose values are updated on faster timescales (e.g. perception of basic visual features of an object), and parameters whose values are updated on slower timescales (e.g. perception and recognition of whole objects).
4. Components of the generative process: finally, we need to determine the components of the generative *process* that the brain/mind is trying to infer based on its generative *model*, and how the generative process relates or is different from the elements composing the generative model.

By considering these factors we can formulate algorithmic theories of different cognitive capacities.

For instance, let us consider human decision-making abilities that enable choosing among the available options. Our Markov blanket of a person whose capacities we are modelling will determine the internal and external states, and their boundaries. Decision-making capacities and behavioral tasks that are used to elicit them are often modeled by partially observable Markov decision processes (POMDP) (see Smith, Friston, et al., 2022). The structure of POMDP presupposes discrete time steps and state variables. Consequently, such models entail that beliefs composing generative models will be represented with categorical variables and parameters. Given such POMDP, they will involve beliefs about observational outcomes, states, actions, and policies. Moreover, depending on the behavioral task being modeled, a generative model can involve different hierarchies of beliefs whose function is to successfully predict sensory observations and enjoy actions that produce preferred outcomes.[20]

5.2.3. The level of implementation

At the implementational level, active inference allows for studying the brain at different levels of granularity. At the level of brain mechanisms, the active inference framework accounts for many predictive loops that characterize different brain systems and its general capacity for regulating internal processes and external environments (see Pezzulo et al., 2022). From a finer-grained perspective, specific components of generative models underlying perceptual and active inferences are related with specific types of neuron populations and different aspects of neural functioning (for a recent experimental study, see Isomura et al., 2023). For instance, studies of the implementations

[20] It is worth noting that process theories of perceptual and active inference are often based on the predictive coding schemes that involve continuous variables and are thought to provide a biologically plausible account of how the brain implements the FEP (see, e.g. Buckley et al., 2017, p. 58). According to predictive coding, perception, learning, and action involve models that minimize prediction errors. Moreover, predictive coding serves as the foundation for the development of more comprehensive predictive processing theories of the mind/brain (Clark, 2016; Hohwy, 2013). In this regard, some consider predictive coding as more than a mere process theory that also offers a computational-level explanation of the brain (Sprevak & Smith, 2023). For more on the relation between predictive processing accounts and the active inference framework, see Hohwy (2013, chs. 2 and 4) and Sprevak and Smith (2023).

of priors and likelihoods of hierarchical perceptual models typically associate them with the functioning of deep and superficial pyramidal neurons, while learning and active inference are typically associated with the activity of dopaminergic neurons (for a short overview and references, see Smith, Friston, et al., 2022, section 5).

In summary, according to the active inference framework, perception, learning, and action are conceptualized as distinct abilities that optimize the same function: the minimization of surprisal by reducing free energy. The generality of the active inference framework allows us to speculate about the holistic function of the mind and its implementation within various neural and bodily processes. Crucially, the generality of the framework also allows us to appreciate the value and feasibility of employing Marr's levels of analysis to comprehensively study the mind.

6. Limits of active inference and the Marrian paradigm

The active inference framework is very ambitious, with its proponents holding very optimistic views regarding its explanatory capabilities. To appropriate a quote from Jakob Hohwy, we might say that this framework purports to

> [G]iv[e] the organizing principle for brain function as such. It should then encompass and illuminate all aspects of perception and action, including aspects that cognitive science and philosophy of mind view as problematic or poorly understood. (Hohwy, 2013, p. 101)

Naturally, due to its broad ambitions, various facets of the framework have been criticized (see, e.g. Bruineberg et al., 2022; Colombo & Wright, 2017, 2018). In this regard, those who are skeptical of the scope or general plausibility of the framework might question its implications for our discussion of Marr's levels of analysis. In what follows, I explain why I think this skepticism does not significantly impact the direction of the present discussion.[21]

The criticisms span a spectrum, with two extremes concerning the generality and empirical adequacy of the framework. On one end, the FEP is often presented as a mathematical principle intended to provide an *a priori* account of formal conditions for modelling self-organizing systems (see Andrews, 2021). When FEP is understood in this way, it is not empirically falsifiable. Moreover, under this interpretation, applying Marr's levels of analysis would need to be reconsidered, because if the FEP lacked empirical constraints for examining mental processes, it would fail to meet the conditions required for analysis at Marr's algorithmic and implementational levels.

However, as discussed in the previous section, even though FEP can be understood as providing a high-level, computational task analysis applicable to all self-organizing agents, its empirical value lies in guiding the development of process theories that aim to model mechanisms, or at least provide sketches of mechanisms underlying various cognitive abilities. This is achieved by constructing generative models with different types of variables and parameters that can be used to test and account for the available cognitive and behavioral data. However, understanding FEP and active inference in more substantial terms leads to another set of concerns.

[21] Thanks to an anonymous reviewer for prompting me to address more explicitly the role of active inference in examining the scope of Marr's levels of analysis.

When active inference and FEP are understood as offering an empirical theory of biological self-organization or mind/brain function, concerns can be raised about its empirical adequacy. For instance, Matteo Colombo and Cory Wright (2017) argue that mental processes involving reward and motivation, which are underpinned by dopaminergic neurotransmitters, are better explained through a pluralistic approach that incorporates diverse theories rather than relying on a single unifying framework. In this view, FEP-based theories that interpret all cognitive mechanisms as involving the minimization of free energy risk oversimplification, by blurring important distinctions across different biological systems, and making it difficult to validate specific mechanisms underlying adaptive behavior.

Such criticisms of the active inference framework, while important, are less relevant to the present discussion. Our focus is on whether Marr's levels of analysis can serve as a framework for understanding the mind as a whole. The purpose of using the FEP and active inference as a theory of the mind/brain is to illustrate (1) what a scientifically influential account of the mind as a whole might look like, and (2) how such an account can be meaningfully analyzed from a Marrian perspective. Admittedly, active inference may not ultimately provide a sufficiently adequate account to explain all cognitive phenomena and every aspect of brain function. However, considering such a general account of cognitive/brain function allows us to explore in a straightforward way what would be required to view the mind as a regulator of cognitive agents and to apply Marr's levels of analysis to the mind understood in this way. Given these considerations, I maintain that Bermúdez has not provided sufficient grounds for thinking that the Marrian framework cannot be applied to the mind as a whole.

6.1. REVISITING THE INTERFACE PROBLEM WITHIN ACTIVE INFERENCE

Nonetheless, it might still appear that the initial problem Bermúdez pointed out has not been resolved. Recall that Bermúdez (2005, ch. 2) discusses the limitations of the Marrian levels of analysis within the context of the interface problem, i.e. the problem of explaining the relation between the personal and different levels of subpersonal explanations. It might still be unclear how Marr's levels of analysis and the active inference framework can be used to address the interface problem. I maintain that expanding upon the previous analysis of the active inference framework through the lens of Marr's levels should dispel such worries.

The personal level, as it is standardly construed, involves thinking about the mind through the lenses of our folk-psychological abilities. At that level, we are discussing agents that possess beliefs, desires, different forms of consciousness, personality traits, and perform actions within their environments (see, e.g. Chappell, 2023; Westfall, 2022). The subpersonal levels are typically construed as, among other things, providing vertical explanations, i.e. explanations about how the personal level abilities are constituted by cognitive capacities as they are studied in psychology and cognitive (neuro)science (see, e.g. Drayson, 2012). Given this standard view, it is easy to see how to apply Marrian levels of analysis to think about the relation between the personal and the subpersonal. Personal level is that at which we detect higher-order psychological abilities that enable agents to act in their environments. This level is, thus, appropriately captured by Marr's computational analysis, which addresses questions about the function a psychological capacity

performs. Given that subpersonal levels, among other things, provide vertical explanations of *how* personal level abilities are formed and constituted, this level can be analyzed in terms of algorithms and physical implementations. Of course, the feasibility of connecting personal and subpersonal levels via Marr's analysis requires the ability to computationally analyze the personal level. However, the viability of accomplishing this was demonstrated by leveraging the active inference framework. This framework posits that the mind operates by minimizing free energy and additionally provides a clear view on how the mind can be investigated from algorithmic and implementational standpoints (see also Sprevak & Smith, 2023).

A further concern might be that the active inference framework cannot provide an adequate account of personal level states and processes. Specifically, some worry that active inference cannot capture the standard roles that conative states, such as desires, play in explaining action and other cognitive phenomena (see, e.g. Klein, 2018; Jurjako, 2022). Indeed, in active inference, action is cast in terms of prior beliefs about policies and preferred outcomes of behaviors. Given this, the equations of active inference eschew mentioning of rewards, desires, and goals, and explain action in terms of prior beliefs (see, e.g. Hohwy, 2013, p. 89; Parr et al., 2022, p. 84; Smith, Ramstead, et al., 2022). Thus, the general worry is that by casting cognition in terms of free energy minimization and Bayesian beliefs, the framework lacks the resources for capturing mental phenomena that involve purely motivational or conative states as they are standardly understood in folk-psychological accounts of agency (for discussion, see Clark, 2020).

However, to think this would involve a confusion between the mathematical description of cognitive processes offered by active inference and a (folk-)psychological description of the same processes. The fact that the FEP equations do not employ terms mentioning desires, preferences or goals, does not mean that they do not refer to functional roles that at the personal level we associate with desires, preferences or goals. As compellingly shown by Ryan Smith and colleagues (2022, p. 8), once we properly distinguish the mathematical level of description from the psychological interpretations of the relevant equations, it should become clear how active inference is compatible and even illuminates some aspects of motivational personal states and processes. To show this, consider how decision-making is typically modeled within the active inference framework.

Active inference models decision-making and planning in terms of minimization of *expected* free energy, usually denoted by G. This is because variational free energy F (see above Equation 4) depends on past and present sensory observations, while to model full-blown decision-making and planning we must also consider future states and observations that actions will produce. Here, decision-making and subsequent actions are understood as processes of inferring the optimal policy that satisfy prior beliefs (e.g. Smith, Friston, et al., 2022). In the context of future-oriented actions, prior beliefs will additionally encompass beliefs about the preferred states and outcomes, determined by an organism's phenotype or preferences. In this regard, expected free energy minimization will involve generative models based on which inferences about optimal policies will be performed. Accordingly, action selection is understood as an inference about which behaviors and policies will produce Bayesian beliefs about states of the world that the organism prefers to observe via its sensorium.

Similarly to variational formulations, expected free energy can also be expressed in several equivalent ways. A structurally analogous equation to (4), goes as follows (the equation is adopted from Parr et al., 2022, pp. 33, 72):

$$G(\pi) = -\underbrace{\mathrm{E}_{Q(\tilde{h},\tilde{s}|\pi)}\,[DKL[Q(\tilde{h}|\tilde{s},\pi)\,||\,Q(\tilde{h}|\pi)]]}_{\textit{Information gain}} - \underbrace{\mathrm{E}_{Q(\tilde{s}|\pi)}\,[\log P(\tilde{s}|C)]}_{\textit{Pragmatic value}} \qquad (5)$$

Here π denotes a policy, i.e. a sequence of actions an agent can choose. $\tilde{s}$ denotes a sequence of sensory observations that is received by an agent over time, while C denotes a set of parameters that encode the agent's preferences about various possible outcomes. $\tilde{h}$ denotes a sequence of hypothesized external states that evolve over time and produce or will produce the sequence of observations. In the first term, $\mathrm{E}_{Q(\tilde{h},\tilde{s}|\pi)}$ indicates the expected value of a distribution, where the distribution $Q(\tilde{h},\tilde{s}|\pi)$ is used for evaluating the value of the distribution $DKL[Q(\tilde{h}|\tilde{s},\pi)\,||\,Q(\tilde{h}|\pi)]$, while in the second term, $\mathrm{E}_{Q(\tilde{s}|\pi)}$ indicates the expected value of a distribution $Q(\tilde{s}|\pi)$ that is used for evaluating $P(\tilde{s}|C)$, i.e. how expected are observations $\tilde{s}$ given a set of preferences C. As before Q denotes a probability distribution that approximates the true posterior probability. The first term measures the expected, i.e. weighted average of the Kulback-Leibler difference between the agent's beliefs about external states given the observed sensory data and the policy ($Q(\tilde{h}|\tilde{s},\pi)$) and the beliefs about external states only given the policy ($Q(\tilde{h}|\pi)$). Minimizing this term encourages the agent to bring its beliefs about external (hypothesized) states in line with the expected external states under the chosen policy. Importantly for our discussion is that the functional role captured by the first term can be associated with commonsensically construed belief-like states whose role is to represent the states of the environment.

In contrast, the second term $\mathrm{E}_{Q(\tilde{s}|\pi)}\,[\log P(\tilde{s}|C)]$ represents the expected likelihood of observing sensory data ($\tilde{s}$) given a set of preferences (C). Minimizing this term encourages the agent to perform actions that will produce expected observations $\tilde{s}$. Accordingly, this term can be associated with desire-like states whose role is to bring about external states that would produce preferred observational outcomes (see Parr et al., 2022, pp. 73-74; Smith, Ramstead, et al., 2022, p. 81). This provides the sense in which (prior) beliefs in the active inference framework, besides personal cognitive states, can also represent notions that refer to goal states, such as ends, values, desires, wants, and so on. Crucially, prior beliefs do not eliminate these motivational states; instead, they encode them through their distinct functional roles in perceptual and active inferences. Thus, despite these states being called beliefs, depending on their role in the relevant equations, they can be thought of as implementing personal level beliefs, desires or other psychological states, traits, and associated processes (for other examples and illuminating discussion, see Smith, Ramstead, et al., 2022).

To further illustrate how such equations can be used to model higher-level psychological abilities, we can consider that equation (5) can be interpreted as giving a free energy formulation of the exploration and exploitation trade-off. The exploration-exploitation trade-off refers to the balance that must be struck when making every-day decisions between exploring new options and exploiting known ones to maximize rewards or values of outcomes. For example, when choosing a

restaurant for dinner, exploring a new one might be a risky choice, but potentially leading to a great new culinary experience. On the other hand, sticking to familiar restaurants typically guarantees a consistently good meal.

In terms of equation (5), this aspect of decision-making can be understood as follows. The left term is often called information gain. It determines the explorative behavior, i.e. how much the organism will seek out new information to resolve uncertainty (i.e. reduce surprisal) about state variables. In our example, this involves determining how much a person would be prepared to try out a new restaurant to learn something new (i.e. what the culinary experience would be like). In this regard, the probability distributions in the left term could also be considered as encoding epistemic curiosity, as it determines the preparedness of an agent to resolve uncertainty about some situation regardless of its current pragmatic value. The right term is called pragmatic value since its probability distributions and parameters determine what kind of observations an organism expects to receive. Thus, it can be understood as encoding different types of goals, including folk-psychological constructs such as different types of conative states (see Parr et al., 2022, pp. 73-74; Smith, Ramstead, et al., 2022, p. 81). In general, agents can minimize their expected free energy by adopting exploratory behaviors, driven by relatively higher values of information gain thereby satisfying their epistemic needs for resolving uncertainty about some states of affairs. Alternatively, they have the option to engage in more exploitative actions, guided by increased pragmatic values that encode preferences for achieving desired outcomes.

These considerations, arguably, indicate that the active inference framework has resources to capture cognitive states, processes, and abilities as they are construed in folk-psychological explanations. This also suggests that active inference can offer a robust framework for understanding cognitive processes at the personal level and that viewing it from the prism of Marr's levels of analysis can further illuminate the relationship between personal and subpersonal levels of explanation.

7. Conclusion

In this paper, I have explored whether Marr's levels of analysis can be used as a general framework for understanding and explaining cognitive phenomena. In a discussion of this issue that has received insufficient attention, Bermúdez argues that many psychological capacities lack algorithmically implementable functions, limiting the utility of Marr's levels of analysis. Moreover, he argues that even if particular non-modular capacities could be analyzed through Marr's levels, this approach cannot be applied to the mind as a whole. To counter these arguments, I offered an example from cognitive science showing how non-modular capacities can be clarified through Marr's levels and drew on the active inference framework, which posits that the primary function of cognition is to minimize free energy. This framework enabled me to articulate the general function of the mind and consider how it can be algorithmically implemented by physical processes. I conclude that Bermúdez has not offered principled reasons for thinking that Marr's levels of analysis cannot be used for understanding the mind as a whole.

Acknowledgments

Initial versions of this paper have been presented at the 30th SIUCC conference in Valencia (Spain) that was dedicated to José L. Bermúdez's work. I wish to thank the participants, especially José, for offering valuable comments on the early version of the paper. Thanks also to Victor Verdejo and Matheus Valente, and others for organizing such a great event. Other versions of this paper were presented at the 11th ECAP in Vienna and the TIPPS Workshop 1 in Rijeka. I thank the participants for their helpful comments. Special thanks go to Stephen Mann for reading and providing constructive comments on several versions of this paper. Finally, I would like to thank the two anonymous reviewers for *Theoria* for their critical and constructive feedback on earlier versions of this paper.

The final revisions of this paper were completed during my time as a visiting fellow at the Faculty of Philosophy, University of Cambridge. I am grateful to Professor Richard Holton for his generous hospitality and support. This paper is an outcome of the project TIPPS (https://sway.cloud.microsoft/xtU0m67IbRR0LAiJ?ref=Link), funded by the Croatian Science Foundation under grant HRZZ-IP-2022-10-1788. My work is also supported by the University of Rijeka under grant uniri-iskusni-human-23-14.

References

Andrews, K., Spaulding, S., & Westra, E. (2021). Introduction to folk psychology: pluralistic approaches. *Synthese*, *199*(1-2), 1685-1700. https://doi.org/10.1007/s11229-020-02837-3

Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology & Philosophy*, *36*(3), 30. https://doi.org/10.1007/s10539-021-09807-0

Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as Marr's computational level: four case studies. *Topics in Cognitive Science*, *7*(2), 287-298. https://doi.org/10.1111/tops.12125

Bechtel, W., & Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science*, *7*(2), 312-322. https://doi.org/10.1111/tops.12141

Bermúdez, J. L. (2000). Personal and sub-personal; a difference without a distinction. *Philosophical Explorations*, *3*(1), 63-82. https://doi.org/10.1080/13869790008520981

Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. Routledge.

Bermúdez, J. L. (2011). *Decision theory and rationality*. Oxford University Press.

Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind*. Cambridge University Press. https://doi.org/10.1017/CBO9781107279889

Bermúdez, J. L. (2020). *Frame it again: New tools for rational decision-making*. Cambridge University Press. https://doi.org/10.1017/9781108131827

Bermúdez, J. L. (2022). *Cognitive science: An introduction to the science of the mind* (4th ed.). Cambridge University Press. https://doi.org/10.1017/9781009064880

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new Markov blankets. *Behavioral and Brain Sciences*, *45*, e183. https://doi.org/10.1017/S0140525X21002351

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *Journal of Mathematical Psychology*, *81*, 55-79. https://doi.org/10.1016/j.jmp.2017.09.004

Chappell, S.-G. (2023). Is consciousness gendered? *European Journal of Analytic Philosophy*, *19*(1), SI8-13. https://doi.org/10.31820/ejap.19.1.7

Chirimuuta, M. (2024). From analogies to levels of abstraction in cognitive neuroscience. In K. Robertson & A. Wilson (Eds.). *Levels of explanation* (pp. 200-221). Oxford University Press.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, *98*, 1-15. https://doi.org/10.1080/00048402.2019.1602661

Colombo, M., & Knauff, M. (2020). Editors' review and introduction: levels of explanation in cognitive science: from molecules to culture. *Topics in Cognitive Science*, *12*(4), 1224-1240. https://doi.org/10.1111/tops.12503

Colombo, M., & Wright, C. (2017). Explanatory pluralism: an unrewarding prediction error for free energy theorists. *Brain and Cognition*, *112*, 3-12.

Colombo, M., & Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*. https://doi.org/10.1007/s11229-018-01932-w

Cooper, R. P., & Peebles, D. (2015). Beyond single-level accounts: the role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, *7*(2), 243-258. https://doi.org/10.1111/tops.12132

Dänzer, L. (2023). The personal/subpersonal distinction revisited: towards an explication. *Philosophy*, *98*(4), 507-536. https://doi.org/10.1017/S0031819123000220

Davidson, D. (2001). Mental events. In his *Essays on actions and events* (pp. 207-224). Oxford University Press.

Dennett, D. C. (1969). *Content and consciousness*. Routledge.

Dennett, D. C. (1981). True believers: the intentional strategy and why it works. In A. F. Heath (Ed.). *Scientific explanation: Papers based on Herbert Spencer Lectures given in the University of Oxford* (pp. 150-167). Clarendon Press.

Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In T. K. Metzinger & W. Wiese (Eds.). *Philosophy and predictive processing* (pp. 148-160). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573109

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415-434. https://doi.org/10.1016/j.neuron.2012.01.010

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, *26*(1), 1-18. https://doi.org/10.1111/phpe.12014

Drayson, Z. (2017). Modularity and the predictive mind. In T. K. Metzinger & W. Wiese (Eds.). *Philosophy and predictive processing* (Vol. 12). MIND Group. http://www.predictive-mind.net/DOI?isbn=9783958573130

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, *28*(2), 97-115. https://doi.org/10.1007/BF00485230

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT Press.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, *29*(1), 1-49. https://doi.org/10.1162/NECO_a_00912

Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience & Biobehavioral Reviews*, *125*, 88-97. https://doi.org/10.1016/j.neubiorev.2021.02.003

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hornsby, J. (2000). Personal and sub-personal: a defence of Dennett's early distinction. *Philosophical Explorations*, *3*(1), 6-24. https://doi.org/10.1080/13869790008520978

Isomura, T., Kotani, K., Jimbo, Y., & Friston, K. (2023). Experimental validation of the free-energy principle with in vitro neural networks. *Nature Communications*, *14*(1), 4547. https://doi.org/10.1038/s41467-023-40141-z

Jurjako, M. (2022). Can predictive processing explain self-deception? *Synthese*, *200*(4), 303. https://doi.org/10.1007/s11229-022-03797-6

Jurjako, M. (2023). Uloga Marrovih razina objašnjenja u kognitivnim znanostima. *Nova Prisutnost*, *XXI*(2), 451-466. https://doi.org/10.31192/np.21.2.13

Jurjako, M. (2024). Are mental dysfunctions autonomous from brain dysfunctions? A perspective from the personal/subpersonal distinction. *Discover Mental Health*, *4*(1), 62. https://doi.org/10.1007/s44192-024-00117-x

Kirchhoff, M. D., Kiverstein, J., & Robertson, I. (2022). The literalist fallacy and the free energy principle: model-building, scientific realism, and instrumentalism. *The British Journal for the Philosophy of Science*, 720861. https://doi.org/10.1086/720861

Kiverstein, J., & Sims, M. (2021). Is free-energy minimisation the mark of the cognitive? *Biology & Philosophy*, *36*(2), 25. https://doi.org/10.1007/s10539-021-09788-0

Klein, C. (2018). What do predictive coders want? *Synthese*, *195*(6), 2541-2557. https://doi.org/10.1007/s11229-016-1250-6

Mallory, F. (2024). Generative linguistics and the computational level. *Croatian Journal of Philosophy*, *24*(71), 195-218. https://doi.org/10.52685/cjp.24.71.5

Mann, S. F., Pain, R., & Kirchhoff, M. D. (2022). Free energy: a user's guide. *Biology & Philosophy*, *37*(4), 33. https://doi.org/10.1007/s10539-022-09864-z

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Murphy, G. L. (2019). On Fodor's first law of the nonexistence of cognitive science. *Cognitive Science*, *43*(5), e12735. https://doi.org/10.1111/cogs.12735

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. The MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Peebles, D., & Cooper, R. P. (2015). Thirty years after Marr's vision: levels of analysis in cognitive science. *Topics in Cognitive Science*, *7*(2), 187-190. https://doi.org/10.1111/tops.12137

Pezzulo, G., Parr, T., & Friston, K. (2022). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200531. https://doi.org/10.1098/rstb.2020.0531

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, *41*(9), 1017-1023. https://doi.org/10.1068/p7299

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889. https://doi.org/10.3390/e22080889

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, *77*(4), 477-500. https://doi.org/10.1086/656005

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press. https://doi.org/10.1093/oso/9780198812883.001.0001

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, *107*, 102632. https://doi.org/10.1016/j.jmp.2021.102632

Smith, R., Ramstead, M. J. D., & Kiefer, A. (2022). Active inference models do not contradict folk psychology. *Synthese*, *200*(2), 81. https://doi.org/10.1007/s11229-022-03480-w

Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science*, tops.12704. https://doi.org/10.1111/tops.12704

Verdejo, V. M., & Quesada, D. (2011). Levels of explanation vindicated. *Review of Philosophy and Psychology*, *2*(1), 77-88. https://doi.org/10.1007/s13164-010-0041-0

Westfall, M. (2022). Constructing persons: on the personal–subpersonal distinction. *Philosophical Psychology*, 1-30. https://doi.org/10.1080/09515089.2022.2096431

Zednik, C. (2018). Mechanisms in cognitive science. In S. Glennan & McKay Illari (Eds.). *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 389-400). Routledge.

MARKO JURJAKO is an Associate Research Professor at the University of Rijeka. During the 2024/2025 academic year, he is a visiting scholar at the Faculty of Philosophy, University of Cambridge, UK. His current research primarily focuses on the philosophy of cognitive science, philosophy of psychiatry, and philosophy of law. He is the author of *Normative Reasons from a Naturalistic Point of View* (FFRI Press). His work has appeared in journals such as *Erkenntnis*, *Perspectives on Psychological Science*, *Biology & Philosophy*, *Synthese*, and the *International Journal of Psychiatry and Law*. He currently serves as Editor-in-Chief of the *European Journal of Analytic Philosophy* and President of the Croatian Society for Analytic Philosophy.

ADDRESS: University of Rijeka, Faculty of Humanities and Social Sciences, Department of Philosophy and Division of Cognitive Sciences, Sveucilisna avenija 4, 51000 Rijeka, Croatia.
E-mail: mjurjako@ffri.uniri.hr – ORCID: https://orcid.org/0000-0002-7252-8627