The private life of the brain: issues and promises in the neuroscientific study of internal states

Héloïse Athéa, Nicolas Heck, Denis Forest

Please cite final version: Athéa, H., Heck, N. & Forest, D. The private life of the brain: issues and promises in the neuroscientific study of internal states. *Synthese* **204**, 64 (2024). https://doi.org/10.1007/s11229-024-04717-6

The private life of the brain: issues and promises in the neuroscientific study of internal states

Héloïse Athéa[1,2], Nicolas Heck[1], Denis Forest[2]

1. Sorbonne Université, CNRS, INSERM, Neuroscience Paris Seine, Institut de Biologie Paris Seine, Paris, France

2. Université Paris 1 Panthéon-Sorbonne, Institut d'Histoire et de Philosophie des Sciences et des Techniques, Paris, France

Abstract

Understanding the relations between mind, brain and behavior is an enduring challenge for science and philosophy. The present article focuses on the concept of internal states, of which emotions are the most studied subtype, in the recent neuroscientific literature. Internal states are conceived by neuroscientists as functional states implemented in neural circuits that drive behavior. To begin, we discuss current definitions of internal states, that emphasize both their intrinsic and relational properties. We stress the difference between preliminary characterizations of these states (that allow researchers to track them) and findings related to their intrinsic nature. We analyze three experimental studies interpreted within this framework, and outline the role of innovative methods in the process of discovery. We interpret the families of states under investigation as homeostatic property clusters, and suggest that the work of Anderson and Adolphs offers a solution to the problem of what constitutes a natural kind in neuroscience. Concerning the relationship between physical states of the brain and mental states, we make explicit the discrepancy between an eliminativist and a reductive project within the literature, and underline the importance of a choice between them. Finally, we suggest that studies of internal states have three interrelated objectives: a better grasp of brain-behavior relationships, a more principled attribution of mental states to nonhuman animals, and better animal models of our own internal states in clinical contexts. With this plurality of objectives comes a plurality of possible outcomes of ongoing research.

Keywords: internal states; brain states; central states; neural bases of emotions; animal models; philosophy of neuroscience

Statements and declarations

Acknowledgments

This work was supported by the Mission pour les Initiatives Transverses et Interdisciplinaires of the Centre National de la Recherche Scientifique and the PhD program "Frontières de l'Innovation en Recherche et Education – Programme Bettencourt". We thank Katleen Pinchaud for her help on the figure. The figure was created with BioRender.

Competing interests

The authors report no competing interests or conflict of interest.

Please cite final version: Athéa, H., Heck, N. & Forest, D. The private life of the brain: issues and promises in the neuroscientific study of internal states. *Synthese* **204**, 64 (2024). https://doi.org/10.1007/s11229-024-04717-6

1. Introduction

There is an emerging field within basic neuroscience: the study of so-called "internal states" or "internal, central states" of the brain (Anderson, 2016), also referred to as "complex behavioral states" (Andalman et al., 2019)[1]. This concept of internal states has been conceived as a theoretical framework to study emotions, defined as functional states implemented in neural circuits that drive behaviors (Adolphs & Andler, 2018; Anderson & Adolphs, 2014; Adolphs, 2018). Whilst different attempts to build a scientific agenda for the study of emotions have been proposed (Pessoa, 2018; Scarantino, 2012), we focus in this article on the above mentioned concept of internal states. Importantly, as internal states are associated with a wide range of behaviors (e.g. exploration of the environment or courtship behavior), the most central goal of internal states studies is to explain patterns of behaviors and transitions from one type of behavior to another; in other words, how appraisals of both external and internal circumstances are encoded in order to shape one's decisions. This literature includes both theoretical work and experimental studies that attempt to define what internal states are, enumerate their properties, but also track their neural signature in organisms such as fruit flies, zebrafish, and mice. Notably, this investigation is driven by recent technological advances, that make it possible to record and analyze neural activity and behavior with unprecedented precision.

Leaving aside more inclusive views of internal states[2], this paper focuses on attempts to produce a workable, useful definition of what an internal state can be in the field of behavioral and affective neuroscience, and related experimental studies. Taking a reflexive stance informed by both philosophy of science and philosophy of mind, our aim is, first, to clarify what an internal state is taken to be, and what the associated view of the relations between brain, mind, and behavior is. It is also to make explicit what questions we shall be better able to address if we extend our knowledge of internal states. Some of these questions are purely theoretical, dealing with the explanation of behavior and with the benefits of the development of neuroscience for scientific psychology. But others have to do with the consequences of a better understanding of affective and motivational states in clinical contexts: identifying how central states shape behavior may be a central goal of animal

1 More on behavioral states and behavioral dispositions in section 3.

2 An example of a (more) inclusive view is Kanwal et al. 2021. The authors are using "internal state" to refer to "the set of cellular, metabolic, and systems level activities that modify how sensory information is dynamically represented and communicated between the body and the brain." (p. 868) Understanding internal states in this way, they encompass all the multidirectional body-brain communications loops studied by physiology, including endocrine mechanisms: in short, they conceive internal states as states of the body linked to regulation and homeostasis. In this paper, we focus instead on another research agenda where internal states are circumscribed to brain-behavior interactions, and related to shifts from one pattern of behavior to another.

models of psychiatric diseases[3]. Our proposal is twofold. First, if we adopt Boyd's view of natural kinds as homeostatic cluster properties (Boyd, 1999), internal states can be conceived as natural kinds. In our view, the problem is less the existence of such kinds than the interest they might have in relation to our scientific and pragmatic goals. From its earliest beginnings (Mill, 1843), the literature on natural kinds has dealt mostly with *kinds of things* (chemical elements, biological species, etc.). Here we have to deal with the question of *kinds of states*: in particular, what constitutes a legitimate kind of state or activity in a special science like neuroscience, and how to connect neural kinds and mental kinds. Second, with the plurality of agendas mentioned above comes a plurality of possible outcomes of research. In a given period of time, some promises may be fulfilled, while others are not.

We begin in section 2 with an overview of the theoretical framework in which experimental studies of internal states are embedded. From this theoretical perspective, we present two different attempts to define what internal states are. Importantly, while these definitions are tools that allow the scientist to *track* and *recognize* internal states, they leave open the question of their ultimate "nature". In section 3, we present three emblematic experimental studies belonging to this research program that attest of its *fruitfulness*: in Kuhnian terms, its ability to discover new phenomena and to suggest new connections between already known phenomena (Kuhn, 1973). By analyzing these results in section 4, we reflect on the importance of the methods used, the results obtained, and on the possibility of considering internal states as natural kinds. In section 5, we discuss the relation of internal states to mental states. We emphasize the difference between two perspectives, a reductive and an eliminative one, and the importance of choosing between them in future research. In section 6, we conclude by considering several scenarios for the future, and suggest that even negative results would have interesting implications.

2. Theoretical Framework and Crucial Properties of Internal States

A natural starting point might be a review of recent neurobiological studies that aim to explain some aspect of behavior by means of internal brain states. However, since most of these studies (like the ones we analyze in Section 3) refer to previously published theoretical reviews, we will first present these reviews which aim to provide a framework for the experimental work[4].

---

3 Abbott, 2020, quotes the acting director of the NIMH, Joshua Gordon: "mental illness is essentially the disruption of internal states. They need to be understood". See below the links between the research on social behavior and animal models of pathological aggression (Hoopfer et al., 2015) and between passive coping and animal models of depression (Andalman et al., 2019).

4 The two reviews, Anderson, 2016 and Anderson and Adolphs, 2014, are explicitly cited in landmark studies in the field, such as (Calhoun & Murthy, 2017; Gründemann et al., 2019). Marques does not cite the reviews, but

The four reviews that we focus on and the book that provides a longer synthesis on the topic (Adolphs & Anderson, 2018), superimpose two different definitions of internal states. First, a definition that is *relational* or *functional*, in which internal states both result from and cause other states. Second, a definition that is more *intrinsic*, based on a list of criteria. Anderson points out that definitions of internal states can be theoretical or operational (Anderson, 2016), but in all of these reviews, the line between the two is not a sharp one, as the authors are trying to answer two very different questions simultaneously: what is the *nature* of internal states (as when they speak of their constitutive elements, or "building blocks"), and how do we *recognize* or *track* them.

Definition 1. The Relational Definition: internal states as central states by reference to which we can explain patterns of behavior

The first of these definitions is presented in Anderson & Adolphs, 2014. Focusing on "emotions" (considered a subset of internal states in later work), Anderson and Adolphs propose that emotions are "central, causative states". As "central states" (i.e., states of the central nervous system[5]), they occupy an intermediate position between the stimuli that trigger them and a series of parallel "responses". These responses can be of the following types: "observed behavior", "subjective reports", "psychophysiology", "cognitive changes", and "somatic responses" (see fig. 2b in (Anderson & Adolphs, 2014)). This web of responses corresponds to what is referred to as "global" or "multicomponent coordination" (Adolphs & Anderson, 2018, p. 77-81). We note that this definition is inspired by a functional account of cognitive/mental states in which such states (such as pain) are identified by specific causal relations linking them to input and output conditions, rather than by their constitutive properties (Putnam, 1967).

Two things are worth noting here. First, the causal relation between internal states and observable behavior seems to be more important in motivating ongoing research than the causal relation between internal states and the other types of "responses" that are listed above. In other words, as several articles illustrate, research on internal states (especially how they are implemented in the brain, at different levels) is stimulated mainly by the promise of a

---

Abbott's interview confirms the connection (Abbott, 2020). Hoopfer et al. and Kennedy et al. are from Anderson's lab (Hoopfer et al., 2015; Kennedy et al., 2020). (Andalman et al., 2019) does not cite these reviews, but earlier papers by Deisseroth do.

5 (Anderson & Adolphs, 2014, p. 188). All internal states are central states, but the reverse is not true. See below, section 5.

more complete and accurate explanation of complex patterns of behavior[6], and by the promise of a more complete and accurate explanation of the transition from one of them to another. This point is important if we want to locate current work on internal states within a larger tradition of research inaugurated by the founders of ethology, Lorenz and Tinbergen (Burkhardt, 2005). In his seminal *Study of instinct*, Tinbergen introduced the concept of "intrinsic central nervous factors" (one subkind of "internal factors" alongside with hormones and internal sensory cues) as a key element of the explanation of the spontaneous behavior of organisms, which requires the integration of behavioral science and neurophysiology (Tinbergen, 1951). The alternative between Tinbergen's hierarchical model of behavioral control and the "hydraulic" model of drive presented by Lorenz (Hinde, 1956; see also below, section 4.2) remains an important source of inspiration for the contributors to the debate on internal states (Anderson, 2016; Robson & Li, 2022).

Second, we cannot easily generalize from emotions to other kinds of internal states. Basic emotions are fundamentally differentiated responses to events located in the external environment[7]. Adolphs and Anderson acknowledge some similarity between reflexes and emotions such as disgust, and even suggest that emotions may have evolved out of reflexes (Adolphs & Anderson, 2018). But in the case of other kinds of internal states (motivation, arousal and drive[8]), the behavior of the corresponding organism is mainly the product of its endogenous activity, and internal factors modulate one's response to external events. This means that if we try to identify the functional role of internal states in general, it is their relation to multiple outputs (somatic, behavioral, cognitive), and not to input conditions, that is typical of them in general. This is why the analogy with the functional definition of clocks is useful: when we say that clocks are devices that measure time, this functional definition is sufficient although it makes no reference to input conditions.

To summarize this first, functional characterization, internal states are central states that i) matter especially in the causal explanation of familiar patterns of behavior, and ii) are not reducible to perceptual or memory states.

---

6 For example, "exploration" *versus* "exploitation" (Marques et al., 2020), or related social behaviors such as aggressiveness and courtship (Anderson, 2016).

7 According to the attitudinal theory of emotions (Deonna and Teroni, 2015), types of emotions (e. g., fear versus anger) are types of *attitudes* towards their object that involve a characteristic readiness to act. There is some agreement between this theory and the functional perspective of Adolphs and Anderson, although Deonna and Teroni conceive the bodily phenomenology of a given emotion as central to the explanation of the corresponding disposition to act (for them, emotions are "felt bodily attitudes"), whereas Adolphs and Anderson conceive psychophysiology and somatic responses as mere consequences of the central emotion states.

8 This list of four typical kinds of internal states appears in (Anderson, 2016).

Definition 2. The Intrinsic Definition: Building Blocks of internal States, or a few tools to detect them

In their 2014 review and their 2018 book, Anderson and Adolphs offer a list of what they consider as emotion "primitives", which they also call the "building blocks" of emotion. These are crucial properties of emotion[9] that are essential to scientific research for at least two reasons. First of all, these properties allow us, in principle, to identify the occurrence of an internal, emotional state in a given organism on the basis of a set of objective criteria, without the help of subjective reports. Secondly, these properties are defined in a way that allows us to ascribe these states to organisms of different species, bypassing the obstacle of the anatomical and physiological differences between them. The list offered in 2014 includes *scalability, valence, persistence,* and *generalization* (also called "generalizability" in 2018). It is not meant to be complete and definitive (Adolphs & Anderson, 2018, p. 65), and it is certainly not entirely original. The combination of scalability (which involves degrees of intensity) and valence (positive or negative) is inherited from a dimensional view of emotions in which each affective state can be located in a bidimensional space (Russell, 1980; Barrett et al., 2007). Persistence, which has no absolute definition (see below), is thought to be a key difference with reflex behaviors. Generalization is less an intrinsic property than a specification of the causal relations mentioned above that are specific to internal states. On closer inspection, generalization seems to be at least two different things. It is the ability for an internal state to be triggered in different circumstances. But it is also the ability to have "pleiotropic", multiple, parallel effects on behavior, physiology, and cognition[10]. This is why Flavell et al. in their 2022 review, taking fear as a prototypical internal state, add a fifth property, and distinguish between generalizability (across contexts) and pleiotropy (multiplicity of concurrent effects) (Flavell et al., 2022).

When Anderson defines what he calls *ΠΠ states* (another name for internal states in general) in 2016, he mentions only *persistence* and *scalability*. In their 2018 book, Adolphs and Anderson explicitly express the idea that the four properties of emotion listed above "surely apply to other internal states as well", which they call MAD states (MAD for motivation, arousal, and drive) (Adolphs & Anderson, 2018, p. 141-142). Accordingly, one

---

9 We avoid on purpose to use "features" as an equivalent of "properties". (Adolphs & Anderson, 2018) deliberately distinguish between *building blocks* (more basic and probably more universal) from *features* of emotion. Valence is a building block (according to them, no emotion without valence), while social communication would be more of a feature (Adolphs & Anderson, 2018, p. 62-63). "Properties" in this case is the generic term, and "building blocks" and "features" denote two different species of properties. In the present paper, it is immaterial to speak of building blocks and properties without distinguishing between them, as we do not deal with what they call features.

10 See the pessimistic biases of bees after vigorous shaking (Bateson et al., 2011).

might think that these shared properties, whatever their exact number, unambiguously delineate the domain of internal states. But this is not exactly the case.

To begin with, if internal states are ultimately brain states, and if the "building blocks" listed above are defining properties of internal states, then these properties should literally be properties of brain states *themselves*. But in fact, this is not the case: scalability is defined at the psychological level, as one can be "annoyed, angry, furious, or enraged" (Anderson & Adolphs, 2014, p. 190), and at the behavioral level. The definition of persistence states that there is persistence when *behaviors* "outlast the stimuli that elicit them" (Anderson & Adolphs, 2014, p. 192). Taken literally, this means that persistence is not a feature of the internal state *per se,* but a feature of the corresponding psychological or behavioral outcome. Moreover, it does not make sense to ascribe a "valence" to neural circuits and neurotransmitters[11].

Of course, this does not mean that we cannot try to give a neurobiological meaning to some of these properties. For instance, scalability evokes the amplitude of neuronal activity (e.g. number of activated synapses, amplitude of synaptic potentials, frequency of action potential). More importantly, the whole purpose of the research program is to discover the mechanisms or circuits responsible for the instantiation of persistence or scalability as defined at the psychological or behavioral level. It is an empirical possibility that a persistent behavioral pattern can be explained, in part, by transient, rather than persistent, activations of groups of neurons (e.g., playing a triggering role in the generation of a lasting/persistent behavioral disposition). And as internal states are defined via "building blocks" in the most general terms possible, it is also an empirical possibility that these properties are realized very differently within the brain of different species. The purpose of experimental research is to move from building blocks to mechanisms, and to find out how different these mechanisms are.

In addition, we can wonder if these "building blocks" are necessary and jointly sufficient to define or identify an internal state, or if any internal state will possess *some* of these features (for instance: scalability), but not necessarily all of them in each and every instance, allowing a certain amount of heterogeneity within the class. The second option seems to be closer the truth. For instance, it is plausible that a state of disgust may not outlast the presentation of the stimulus that elicited it: this would be a case of emotion without *persistence*. Moreover, if in the case of emotions it is easy to understand that valence has to

11 Adolphs and Anderson (2018, p. 70) mention the proposal of Edmund Rolls, a neuroscientist whose aim was to replace the subjective view of valence with objective criteria. But these criteria are behavioral (linked to reward and punishment), they are not provided by neuroscience.

do with antithetical pairs (positive or negative), drives do not come in pairs: what would be the positive counterpart of thirst or hunger? With this second, more modest option, we will have interesting *family resemblances* between internal states, each of which combines some of the properties listed above, and striking differences between internal states and other states of the nervous system, rather than a homogeneous category (see below, 4.2).

In summary, the "building blocks" identified in the theoretical reviews are less intrinsic or essential properties of internal brain states than tools used in order to detect them in experimental settings.

3. Experimental studies of Internal States

Now that we have an overview of what "internal state" means in the recent scientific literature, and of the reasons why scientists are interested in making discoveries about internal states, we can focus on the experimental studies that aim to identify these sates in the brain. Recently, several research articles have provided some information about the profile of internal states. The general approach, guided by Anderson and Adolphs' theoretical framework, is to combine behavior and brain activity recordings with experimental manipulations of brain activity in order to unravel which specific neural networks drive behavioral states. Below we present three publications that illustrate the success of this approach. It's noteworthy that these three articles are representative of a large number of experimental studies published in the recent years, and conducted with different animals models including mouse, zebrafish, fruitfly and *C. Elegans* (e.g. Calhoun et al., 2019; Deutsch et al., 2020; Gründemann et al., 2019; Ji et al., 2021; Kennedy et al., 2020; Liu et al., 2022; Lovett-Barron et al., 2017). Interestingly, these studies are the result of a renewal of research methods: new large-scale neuronal recording techniques, new behavioral measurement techniques, and the development of artificial intelligence. The engineering of electrical and optical probes makes it possible to simultaneously records tens of thousands of individual neurons in the rodent brain, or each of all neurons in the nervous system of *C. elegans* and zebrafish larvae, two widely used animal models (Urai et al., 2022). The behavioral observations were improved in two ways. The number of parameters recorded has been multiplied, and each parameter is quantified with high precision. For example, locomotor activity can be decomposed into parameters such as eye movements (allowing identification of the animal's visual field), paws or tail movements, head-direction, and so on. This advance is in line with a recent "ethological" trend in neuroscience, which also consists in carrying out these measurements in the context of spontaneous behavior (Anderson & Perona, 2014; Krakauer et al., 2017). As these techniques have improved, an unprecedented amount of data has been generated that can only be analyzed with tools made available by the

concomitant rise of artificial intelligence, especially deep learning (Calhoun & Murthy, 2017; Datta et al., 2019; Mathis & Mathis, 2020).

Study 1. Marques et al., 2020

In a technical *tour de force*, Marques et al. monitored the activity of every single neuron in the brain of freely behaving zebrafish larvae (Marques et al., 2020). More precisely, they measured the calcium level in each neuron, as increased neuronal activity can be reliably inferred from increased calcium levels. Whilst no prior hypothesis was set on the specific contribution of any particular neuronal activity to the overall behavior, the collected recordings unraveled a correlation between the activity of neurons from specific parts of the brain and different types of specific movements, but also of more generic behavioral choices. The authors were able to localize activated neurons in correlation with specific visual signals and motor actions. (e.g. prey detection in the visual tectum, swim turn in motor networks). These results are similar to those obtained, for example, by recording brain activity with electroencephalography of brain scanner while a subject receives a sensory input or performs an action (albeit here we have single-cell resolution in the whole brain of a freely behaving animal). But the most important finding in this study concerns the transition from one pattern of behavior, or "behavioral state" (exploration) to another (exploitation). In the exploration state, the larva travels long distances, while in the exploitation state, which corresponds to hunting and feeding, other behavioral features (e.g. eye convergence, turns…) are observed. Importantly, the larva oscillates between these two states in a similar manner when fed and when starved, which underlines the importance of internal factors in the switch from one behavioral state to another. The authors then look for the correlates of the exploitation state. On the one hand, they identify, a "trigger network" located in the ventrolateral habenula, dorsal raphe and rhombencephalon, whose activity correlates with the transition from the exploration to the exploitation state. On the other hand, they have also found a specific subpopulation of neurons in the dorsal raphe, whose sustained activity corresponds to the exploitation state itself (they call them "exploitation state encoding neurons" or more simply "exploitation-state neurons") (Figure A1). The authors suggest that this dorsal raphe subpopulation encodes a "generalized motivational state".

Beyond the localization of neuronal activity, Marques et al. have found a correlation between the profile of the neural activity and the duration of the exploitation state. The amplitude of the peak of the neuronal calcium rise in the trigger network (a sudden rise, followed by a linear progressive decay) is mirrored by the peak of the activity of the state-encoding neurons, and both are predictive of the duration of the behavioral state. The

behavioral state ends only when the calcium rise is back to baseline (Figure A2). This is an important result, as properties that had been proposed to define internal states at the psychological and/or behavioral level (see above, section 2) are found to be properties of the corresponding neural activity: scalability (the magnitude of calcium rises correlates with the duration of the state), and persistence (in the state-encoding neurons, the calcium rise is maintained during the behavioral state, which lasts on average 7 minutes).

Study 2. Hoopfer et al., 2015

The 2015 paper by Hoopfer et al. represents another endeavor to understand the relationship between different patterns of *Drosophila* behaviors: its authors (that include David J. Anderson) focus on the relationship between two social behaviors, courtship and aggression in male conspecifics (Hoopfer et al., 2015). They build on previous work that identified a population in the murine hypothalamus whose stimulation is sufficient to trigger aggressive behaviors, but which is also involved in mating behavior. If Marques and collaborators were studying motivation, Hoopfer et al. are tracking not only motivation, but also "social arousal", another subtype of internal states. Social arousal is not just a physiological state of wakefulness, but corresponds to an enhanced sensitivity to social cues.

Previous studies have linked the activity of P1 neurons (a part of the posterior medial protocerebrum in the fruitfly brain) to courtship, but no to aggression behavior. To obtain a more complete functional profile of these neurons, this study relies on optogenetics, a method that allows rapid and reversible modification of the activity of selected neurons at desired time points (Boyden et al., 2005). Activation and inhibition of the P1 neurons by optogenetic manipulation induced and blunted the aggressive behavior, respectively, suggesting their involvement in aggressive behavioral states (Figure B1). However, several other key findings were obtained.

First, the authors established that the aggressive behavior could be induced in the absence of a reaction from another Drosophila, meaning that the aggression induced was a fly-*intrinsic* response and did not depend on counterattack from the other fly. Second, the authors established that different levels of activation of these neurons (scalability, again) were sufficient to trigger aggression and courtship. A weak activation of P1 neurons induced aggressive behavior, while a higher activation induced both aggressive and courtship behavior. Third, the authors demonstrated that the transient stimulation of P1 neurons was sufficient to induce a lasting behavioral disposition. Drosophila were isolated from other flies by a barrier to prevent induction of aggression through social feedback (i.e., counterattack) or

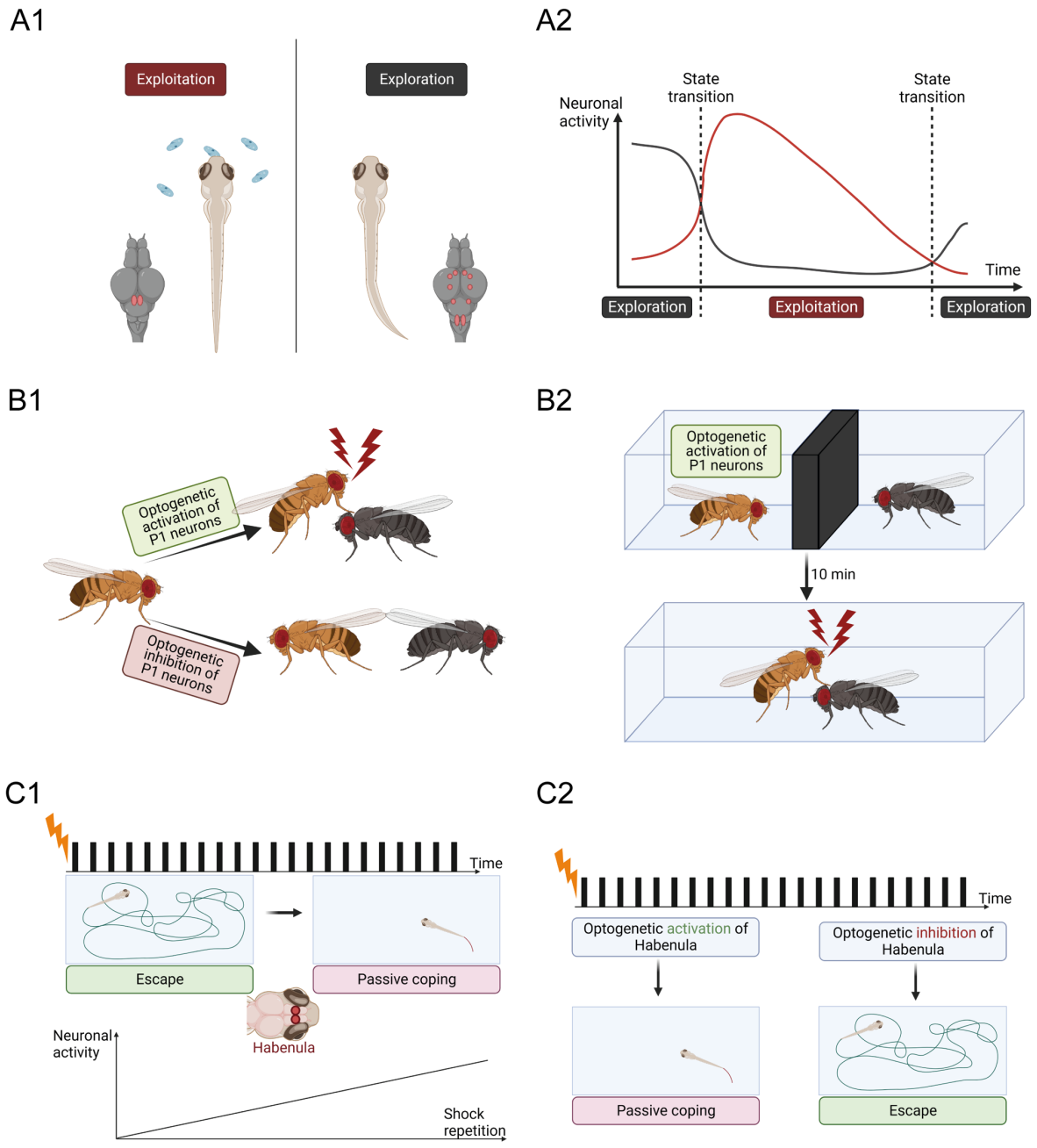
behavioral self-reinforcement. Under these conditions, optogenetic activation of P1 neurons induced courtship behavior of the isolated fly. But, when the barrier separating two male flies was removed, then aggressive behavior was observed even after the optogenetically induced courtship behavior had disappeared for at least five minutes (Figure B2). In other words, the experimental activation of a set of neurons did not only trigger a fast, "overt" response, but also a persistent, "covert" state of arousal, that was sufficient to induce aggression in the presence of another male fly several minutes later[12]. This means that the persistence of internal states is compatible with two different relationships to behavior: the promotion of a lasting *behavioral state* (see above, Marques et al., 2020), and the production of a *lasting disposition* or state of aggressiveness that can persist without overt motor activity.

Study 3. Andalman et al., 2019

A third example is the study by Andalman et al. from Deisseroth's laboratory, who developed a strategy to identify neurons involved in the behavioral state transition from *escape* to *passive coping* (Andalman et al., 2019). Considering that different defensive behaviors can occur depending on previous experience, zebrafish larvae were exposed to a mild shock to which they initially responded with escape behavior. However, unavoidable repetition of the shock led to passive coping. Whole brain recordings of calcium ion levels were performed (as in the Marques study). Andalman et al. identified a progressive increase in calcium levels in the ventral habenula throughout the course of the delivery of repeated mild shocks (Figure C1). Single neuron resolution recordings showed diverse latencies and durations, but all neurons increased progressively their activity. Selective activation of habenular neurons by optogenetics induced passive coping, and the inhibition of the same neurons elicited escape in fish that had previously acquired passive coping behavior due to repeated shocks (Figure C2). From a methodological point of view, this study combines the resources of correlation between behavior and brain activation (as in the study by Marques et al.) and optogenetic manipulation (as in the study by Hoopfer et al.), while applying them to the same target. Their results suggest that the activity of habenular neurons is necessary and sufficient for passive coping, and also that these neurons also play a key role in the transition between the two behavioral strategies.

12 While the persistent aggressive behavior can be induced by the stimulation of P1 neurons, these neurons are only transiently active. As such, P1 neurons only promote the behavioral state. In a subsequent study conducted by the same laboratory, it was demonstrated that a transient activity of P1 neurons induces persistent activity in another set of neurons (PcD neurons), and this latter activity is necessary for the persistence of the behavioral state (Jung et al., 2020).

Please cite final version: Athéa, H., Heck, N. & Forest, D. The private life of the brain: issues and promises in the neuroscientific study of internal states. *Synthese* **204**, 64 (2024). https://doi.org/10.1007/s11229-024-04717-6



Please cite final version: Athéa, H., Heck, N. & Forest, D. The private life of the brain: issues and promises in the neuroscientific study of internal states. *Synthese* **204**, 64 (2024). https://doi.org/10.1007/s11229-024-04717-6

4. Methods, results and ambitions

4.1. From methods to findings

As we have seen, these three innovative studies are the product of recent developments in experimental methods. By this, we do not mean that the development of tools (calcium imaging, optogenetics, and so on) has in itself caused a genuine scientific revolution (as suggested by Bickle, 2016). It is rather that these methods function as enabling conditions for new research. On the one hand, this research is not driven by technological innovation, but by a program with a strong theoretical background (see above, section 2): the concept of internal state is a generalization of the view of Anderson and Adolphs, according to which emotions can be studied as biological phenomena in non-human animals without relying on introspection (Anderson & Adolphs, 2014). On the other hand, the experimental work of Marques, Hoopfer, and others is designed to provide answers to questions that have already been addressed by a large community of researchers over a longer period of time. Such questions are located both at the methodological level - how to collect direct, physiological evidence about the internal factors of behavioral choices; how to combine direct and indirect (behavioral) evidence (Tinbergen, 1951), and at the level of scientific content - what neuroscience has to say about behavioral choices, behavioral dispositions, and states like motivation or drive. Findings made possible by new research tools may lead to new questions, but the current use of these tools provides resources for puzzle solving, not conceptual change (Parker, 2018).

To claim that one has identified "state encoding neurons" in the brain, or a "hidden variable that shapes the temporal structure of motivation and decision making", is to claim that one has identified a causal link between the internal state of the brain and the corresponding behavioral output (e.g., exploitation). The ambition to provide causal explanations is even more obvious in optogenetic studies, which rely on the manipulation of the system to find on which pattern of activation the transition from active to passive coping depends. This leads to classic methodological worries about the adequacy of experimental methods to establish causal relations (Silva et al., 2013 ; Craver 2021). A correlation between activity in the raphe nucleus and the exploitation state is not in itself evidence of causation. Concerning manipulation, one of the advantages of optogenetics is the combination of stimulation and inhibition to demonstrate that the activity of the target of the optogenetic intervention is both necessary and sufficient to produce a given effect. However, the afferent problems are now well known (e.g. see Otchy et al., 2015). Inducing a behavior by activating some neurons does not bring evidence that these neurons are involved in the same behavior in unmanipulated animals, and possible off-target effects of the optogenetic stimulation may contribute to the explanation of its effect.

On the other hand, inhibition does provide evidence that these neurons are a necessary to produce the effect, but many other neurons may also be necessary.

However, on a more promising note, the use of independent experimental procedures, with different organisms, in the context of different behavioral options (exploration versus exploitation, courtship versus fighting, passive coping versus flight) gives a significant degree of robustness (Wimsatt, 1981) to the preliminary conclusions of these studies, based on converging evidence. 1. Changes in brain activation patterns, within restricted parts of the brain, correspond to mutually exclusive types of behavior (more generic and more persistent than specific motor actions), and in certain cases, these changes can allow us to predict the transition from one type to another. 2. Some key properties of internal states, like persistence and scalability, initially defined at the psychological or behavioral level, are expressed at the neural level. 3. Roughly speaking, there is a distinct mode of brain functioning, with a distinct temporal scale, that differs both from the immediacy of reflex action and from the endogenous activity associated in humans with rumination and introspection (as in the activity of the default mode network); it is involved in the evaluation of environmental circumstances and the definition of behavioral strategies.

These encouraging results lead us to a general question about how to measure the success (or failure) of the research program that we are analyzing. In their 2018 book, Adolphs and Anderson commit themselves to a mechanistic view of scientific explanation, in which the accurate description of a causal mechanism allows us to "predict the behavior of a system", "to intervene and manipulate the system to produce specific results", and could also, at least in principle, allow us to "build such a system from scratch" (Adolphs & Anderson, 2018, p. 109). They understand "mechanism" as "units" located at different levels of organization and the causal relationships between them (Adolphs & Anderson, 2018, p. 109)[13]. In the case of internal states, such mechanisms will include key units that play a distinctive role in the production of behavioral states. At this preliminary stage of research, one can think that the description of the mechanism susceptible to explain the occurrence of any instance of internal states is still incomplete: higher-level units (neural circuits) and patterns of activation are identified in the above-mentioned examples, but neuromodulators are not, at least not in a systematic fashion, and the organization of the mechanism is not fully known.

The question we want to raise, however, is not about exhaustivity in the description, but whether there would be an invariant neural signature of internal states. Indeed, the exper-

13 They do not distinguish as clearly as Machamer, Darden and Craver in their classic paper between "entities" and "activities" (Machamer et al., 2000), but they roughly refer to the same class of biological objects.

imental findings represent a first convincing step towards identifying a neurobiological signature of internal state. Robson and Li, who led the study published by Marques et al., have analyzed neural recordings found in the literature and proposed that state-encoding neural activity is characterized by a sudden rise followed by a linear progressive decay, with duration correlated to the duration of the behavioral state. Similarly, Flavell and colleagues (2022) suggest that different internal states are induced by neuronal populations that share the following features: the neurons have brief activity leading to persistent states, and have long-range projections across the brain to control different aspects of the state and its behavioral expression. These authors claim that further methodological improvements and empirical findings will help to define internal states, and envision that they could eventually be described from brain activity alone.

However, when Marques et al. observe that in the zebrafish larva, the profile of activation sustaining exploration, with its typical initial rise, is quite different from the one that sustaining exploitation (Marques et al., 2020, p. 242): it is an important (negative) result, because if research is looking for an invariant neural signature of internal states (or of subcategories, like motivation), such a result raises the probability that there is in fact no such signature waiting to be discovered. The question, then, is about the relationship between variability in the neural realization of internal states, the legitimacy of the category itself, and the prospects for ongoing research. We want to suggest that, in the long run, the interest of the category "internal states" will not depend mainly on the degree of similarity between the underlying mechanisms and on the identification of neural signatures. Rather, it will depend on the benefits of the research on internals states for our epistemic and non-epistemic purposes. We develop this point in the next section, with a reference to the philosophical literature on natural kinds, and in the general conclusion. But the point is also related to our understanding of the relation between, on the one hand, the preliminary characterization of internal states in terms of properties (including generalization and pleiotropy, that is, relational properties) and, on the other hand, their characterization in terms of mechanisms. We suggest that the latter complements the former, but cannot and does not have to replace it. Some categories in neuroscience (like motor cortex, or place cells) are partly relational and cannot be reduced to types of structural properties or intrinsic patterns of activation. We believe that the unity of the category of internal states should also be thought of as a combination of intrinsic features like scalability and persistence, and relational, or functional features. Since variation is expected and should be tolerated in the realization of a property like scalability, functional properties will remain essential to the characterization of each type of state.

4.2. Internal states as natural kinds

In our view, the most adequate philosophical characterization of the working hypothesis that is central to the study of internal states is the following: we should think of internal states in terms of homeostatic property cluster (HPC) natural kinds (Boyd, 1999; Craver, 2009; Griffiths, 1999; Slater, 2015). According to Boyd's view, the key features of such kinds are the following:

(1) The existence of a family of properties that are "clustered in nature" – which means that they co-occur frequently, but not necessarily.
(2) This co-occurrence is the result of a process called 'homeostasis': either the presence of some properties favors the presence of the others; or there are underlying mechanisms or processes that maintain the presence of these properties (or both).
(3) This clustering of properties and their conditions produces effects that we judge important, either theoretically or practically (or both).

Even if internals states are *prima facie* quite different from Boyd's favorite examples of HPC kinds (in particular, biological species) one can point out that Griffiths has already applied Boyd's view of natural kinds to the case of emotions (Griffiths, 2004). We suggest that the view of Adolphs and Anderson can be expressed as follows:

(1) Internal states possess a family of properties that are contingently clustered (e.g., persistence, generalizability and scalability, often co-occur).
(2) The co-occurrence of such a cluster of properties depends on an underlying mechanism. It is possible that this underlying mechanism does not consist exclusively of entities and activities located within the brain (e.g., think of the external factors of emotions), but these internal factors are always present and important.
(3) Internal states, as clusters of homeostatic properties, have physiological and behavioral consequences that are important from a biological point of view and potentially relevant for practical (clinical) purposes. Accordingly, knowledge about internal states is potentially useful for goals like explanation, prediction and control.

As attractive as this characterization of internal states is, it runs the risk of inheriting well-known problems associated with Boyd's view of natural kinds (Craver, 2009; Slater, 2015). In particular, one important requirement (also called "accommodation") expressed by Boyd is that differences between real kinds should mirror differences between corresponding mechanisms (taxonomies should "carve nature at its joints"). But here comes a risk of regress, because the question of what defines a natural kind becomes the question of what defines a (legitimate) kind of mechanism.

Please cite final version: Athéa, H., Heck, N. & Forest, D. The private life of the brain: issues and promises in the neuroscientific study of internal states. *Synthese* **204**, 64 (2024). https://doi.org/10.1007/s11229-024-04717-6

We suggest that an important aspect of the research on internal states is to offer pragmatic criteria for defining what counts as a legitimate kind of neural state: equivalent mechanisms, whatever their differences in composition and organization, are those that are responsible for the same clustering of properties and have the same type of behavioral consequences. This is why one can say that flies and rodents are in the same internal, central state (e. g., the one that promotes aggressiveness), although the details of the implementation may differ in many ways (Anderson, 2016). This is consistent with the view that HPC kinds are multi-realizable.

But sameness and difference, at the neural and/or at the behavioral level, come in degrees, and which degree is important to consider depends on what researchers intend to study and try to achieve. In our view, this means that reference to HPC kinds should be quite flexible to accommodate the various purposes of research, rather than the structure of the world (Craver, 2009). Within Anderson's theoretical project, it is important to think of internal states in general as a HPC natural kind: clustered properties that modulate the behavioral strategies of organisms because of underlying mechanisms. Concerning the homeostatic mechanisms responsible for this kind of states, it is plausible that Lorenz and Tinbergen were both on the right track in the light of recent findings listed above (Anderson, 2016). The scalability of activity within the system (as measured via calcium imaging in Marques et al., 2020) is important to define the persistence of internal states, which is consistent with the Lorenz's hydraulic model; meanwhile, a hierarchy of nodes is an important means to implement nested behavioral decisions (Hoopfer et al., 2015), with explicit reference to Tinbergen. To think that hierarchy is important, however, is not holding that it must be necessary. We simply cannot know in advance what the mechanisms responsible for the occurrence of internal states will look like in a given species that has not been studied yet, and how much of the machinery for internal states will be conserved through evolution. If scalability and persistence are essential to what we have called a distinct mode of brain functioning, it is an empirical possibility that different "recipes" for producing scalable and persistent states coexist in nature.

If, as we have just seen, there are reasons to think of internal states in general as an HPC kind, there may be in parallel reasons to think of internal states associated with emotions as a distinctive HPC kind (with a distinctive causal import), or to think of an even more specific type of internal state (e.g., the one responsible for exploitation, or for passive coping) as a genuine HPC kind. Each type of kind will be grounded in a set of facts. Note that it would be erroneous to think that narrower kinds exclude multiple realizations. At the neural level, for example, there may be few interesting cross-species generalizations about motivation. To

take another example, large differences in the causal explanation of passive coping in the zebrafish and learned helplessness in humans would not undermine the hypothesis that both involve a type-identical internal state (because of the clustering of properties and the similar output), but the interest for psychiatry of the work of Andalman, Deisseroth and colleagues would diminish. The question, then, is less the existence of HPC kinds, than the interest of such kinds for prediction, explanation and clinical research. Even if Adolphs and Anderson are looking for robust, objective categories[14], the ones that shall prevail in the literature will not be more objective, but only more useful than others.

5. Internal states and mental kinds

The enquiry about internal states began as an enquiry about emotions (Adolphs & Anderson, 2018; Anderson & Adolphs, 2014), and our first, pre-scientific characterization of internal states comes from introspective access to our own mental life (Anderson, 2016). In this context, it is inevitable to address the relation between internal states and mental states: after all, any state of the brain can be called "central", but only those that can be connected with familiar psychological categories like emotion and motivation are "internal, central states" (Anderson & Adolphs, 2014, p. 197)[15]. This is consistent with standard linguistic conventions, where the word "internal" refers to covert aspects of mental life and subjectivity, as in the case of internal monologue, also called inner speech (Alderson-Day & Fernyhough, 2015). What remains to be seen, however, is how to properly characterize the mental dimension of internal states in this context.

Adolphs and Anderson both endorse token physicalism (e.g. Adolphs & Anderson, 2018, p. 44): emotions or motivational states are brain states, they could not modulate our behavior if they were not encoded in the brain. In the long run, the task of neuroscience is to discover how. But according to them, *from an epistemic point of view*, it is more fruitful (and in practice, unavoidable) to start with functionalism: in the initial phase of the research, the definition of a kind of mental state has to be functional, not neurobiological. It must be functional for two reasons. First, as we have seen, functional definitions are *relational* definitions (Adolphs & Anderson, 2018, p. 40), in which a kind of state is characterized in terms of causes (input conditions) and effects (physiological and behavioral outputs). In experimental con-

14 On this crucial point, Adolphs and Anderson are not entirely consistent: officially, they embrace scientific realism; science aims at (and gets closer to) an objective description of the world (e. g., Adolphs and Anderson, 2018, p. 293) but they also write "one can choose to taxonomize the states of an organism in many different ways *and they need not be mutually exclusive if they are scientifically useful*" (emphasis added, ibid., p. 61). This latter phrasing justifies to a certain extent our liberal use of the concept of HPC kinds.

15 One could argue that, according to this distinction, the neural correlates of perceptual states or memories should count as internal states, not only emotions, motivational states, arousal and drive. The answer seems to be that Adolphs, Anderson and others think that the internal states that they consider have enough in common (in terms of properties, functional role, and neural characteristics) to constitute a class of their own.

texts, this approach is presumed to be the only one that allows us to fix the reference of terms like "fear", "drive", or, to take the canonical philosophical example, "pain". The second reason is that, since the physical realization of fear or arousal may vary from species to species, or, to put it differently, since we do not know to what extent it varies, functional definitions are the only ones that can be sufficiently general to allow us to ascribe emotional and internal states to organisms belonging to different species.

Still, functional characterizations may take different forms, two of which being judged both important and complementary. To use Adolphs and Andler's terminology, we can distinguish between *causal role* functionalism and *etiological* functionalism (Adolphs & Andler, 2018). *Causal role* functionalism, which is close to Putnam's original view of organisms as probabilistic automata (Putnam, 1967), characterizes psychological states in terms of (typical) proximate causes and behavioral effects. In contrast, *etiological* functionalism is identified with an evolutionary understanding of how a functional state solves an ecological problem (Adolphs & Anderson, 2018). One of the main motivation for adding a "broader" functionalist view to (narrow) causal role functionalism is to be able to distinguish between emotional/internal states that carry their proper function[16], and pathological states that do not subserve such a function, like phobias and harmful varieties of anxiety (Adolphs & Anderson, 2018, p. 48). This is an important point for Adolphs and Anderson, even if the ability of the evolutionary perspective to draw the line between states that carry their normal function and pathological states is debated (Murphy & Woolfolk, 2000; Wakefield, 1992; Faucher & Forest, 2021). One thing remains certain: psychofunctionalism is there to define a framework that allows researchers to study the mental life of animals in a fruitful manner, while remaining agnostic about their own introspective access to their mental life.

In our view, this provisional, methodological functionalism masks important differences both in the possible outcomes of research and in the underlying view of the mental within the field. In terms of future scenarios, Patricia Churchland (Churchland, 1986, p. 284-285) has made an interesting distinction that can be applied to research on internal states, as it concerns the coevolution of psychology and neuroscience. According to her, the reduction of psychological kinds can be either "smooth", or "bumpy". If reduction is smooth, mental kinds will square with neural kinds: research will find something like a neural signature of emotions, or fear (plausibly, as we have seen, with generic, rather than very specific features). If the reduction is bumpy, the taxonomy of internal states will not reflect common sense distinctions between emotions, motivational states, and so on. Adolphs and Anderson explicitly

16 Proper function being what states of the same kind have been selected for – e.g., the function of states of fear would be to trigger an adaptive response in the presence of danger.

consider this second scenario as a genuine possibility, and do not commit themselves to any familiar taxonomy of mental states (Adolphs & Anderson, 2018, p. 152). We simply do not know in advance how transformative science will be, although we can expect familiar taxonomies to have die-hard supporters.

Besides this variety of scenarios for the future, we think there is some disagreement within the field about the meaning of mental concepts. In the literature, there are two mutually exclusive views of internal states in their relation to the mental world. According to the first, one key benefit of research on internal states is to allow us to extend out psychological knowledge, to ascribe specific mental or cognitive states to nonhuman animals, and to discover how these states are realized in the brain. If this view is taken seriously, it is impossible, for instance, to think of "social arousal" as a kind of state without having to deal with its intrinsic "aboutness" (reference to conspecifics). But, according to the second of these views, there is an eliminative element involved in the co-evolution of psychology (or ethology) and neuroscience. To define mental kinds as functional kinds is to begin a process in which we gradually lose our initial interest in psychology. For instance, what we call motivation in the zebrafish larva is simply the "exploitation-encoding-state" that explains the corresponding behavior. Once we have a neurobiological description of the state, there is no need to add any psychological description. The psychology of motivation is eliminated by the progress of neuroscience. Interestingly, the experimental work of Marques et al. (above, section 3) has been supervised by Robson and Li, who advocate a dynamic system view of neuroethology in which there is no room for representations, only for a "neuromodulatory state space" in the brain (Robson & Li, 2022). At best, mental concepts would function as convenient labels, and mental states would subsist as mere kinds of behavioral dispositions, the task of neuroscience being to identify the causal basis that explains the presence and persistence of these dispositions. Each type of internal brain state would be defined as the causal basis of a behavioral kind. In this second option, the difference between physical states and mental states becomes the difference between the causal basis of the behavioral dispositions and the corresponding dispositions themselves.

In our view, in the long run, a choice has to be made between behavioral neuroscience and cognitive/affective neuroscience: either internal states are understood as the physical basis of mere behavioral dispositions, or they have to do with the mental and the cognitive, including what is felt. While the first option is more parsimonious, the second one seems more attuned to what we have called the clinical project, in which the modeling of human mental conditions is central. Within this second option, research on internal states cannot remain forever agnostic about the existence of subjective, conscious states in nonhuman animals, a hot topic for interdisciplinary research today (Birch et al., 2020). If the second option prevails,

research will have to deal with valence, appraisal and pessimistic biases as genuine *explananda*, not as constructs that are there only to be discarded later.

6. Conclusion

One of the main interests of research on internal states is that it deals with fundamental questions in an exemplary manner: how to define a kind of state of the brain/ neural system, how to make progress in neuroethology, and whether it is possible to move beyond the study of neural mechanisms in specific animal species to reach robust cross-species generalizations.

Research as the one described in this paper is a complex affair, in part because it has at least three interrelated goals. The first is to advance our understanding of brain-behavior relations. The second is to allow us to ascribe mental states to nonhuman animals in a principled way. The third is to provide animal models of our own internals states that may be helpful in clinical contexts. These three goals define three interrelated research agendas, and three corresponding scientific perspectives on the same family of internal states, each with more or less emphasis on the mentalistic description of these states, or on the clinical implications of the research.

Our own bet is that research on internal states is promising, at least in the sense that, if we judge its fruitfulness by recent findings, it cannot fail to yield some more results. But the first of these goals seems easier to achieve than the other two, and the second one is by nature exposed to controversy and philosophical disputes. For the time being, we can at least imagine several scenarios for the future, with varying degrees of optimism.

A *bright future*, in the style of Deisseroth (Deisseroth, 2021): more precise animal models of affective and motivational states are developed, making it possible to propose new types of therapeutic interventions in psychiatry (e.g., helping to change behavioral patterns of depressed patients). A *favorable case*, in which the development of the Anderson and Adolphs' research program yields a strong return: progress is made in the understanding of the neural machinery that controls behavior and behavior change; and/or in defining the neural underpinnings of affective and motivational states. We reach robust neurobiological generalizations. From there, perhaps, we could move from the favorable case to the most favorable case mentioned above. But it is also conceivable that generalization would not progress in a uniform manner. For instance, knowledge of the neural circuits involved in aggression in nonhuman animals might not shed much light on human forms of pathological violence, while we make progress on the neurobiological factors of learned helplessness and depression. And it is also conceivable that if we succeed in the identification of internal factors of one's dispositions, the degree of manipulability of these factors will remain very

weak in the current state of our knowledge. Finally, there is the *least favorable case*, where findings remain inconsistent and predictions remain unfulfilled. In this case, even if internal states are genuine natural kinds (as defined in 4.2), their scientific interest is limited. From the point of view of neuroscience, either emotions, motivational states, drives, and arousal do not have enough in common in their realization, and lumping them together does not lead to interesting generalizations. Or the mechanisms associated with each type of internal state vary too much from species to species. In this latter case, we would probably have to focus more on the human brain for clinical purposes. This would be bad news for animal models in psychiatry, but on a brighter side it would also help neuropsychiatric research to move forward and refocus.

When Alison Abbott presented this field of research to the readers of *Nature* in 2020, she called her paper *What animals really think* (Abbott, 2020). In the absence of a crystal ball that would allow us to decide between the three scenarios mentioned above, what we can already say with certainty is that if we do not (yet?) know what animals really think, all this research brilliantly opens the way to a better understanding of their internal states, and possibly of our own ones.

References

Abbott, A. (2020). What animals really think. *Nature*, *584*, 183-185. https://doi.org/10.1038/d41586-020-02822-3

Adolphs, R. (2018). Emotions are functional states that cause feelings and behavior. In A. S. Fox, R. C. Lapate, A. J. Shackman, & R. J. Davidson (Éds.), *The nature of emotion : Fundamental questions*. Oxford University Press.

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion : A new synthesis*. Princeton University Press.

Adolphs, R., & Andler, D. (2018). Investigating Emotions as Functional States Distinct From Feelings. *Emotion Review*, *10*(3), 191-201. https://doi.org/10.1177/1754073918765662

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech : Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*(5), 931-965. https://doi.org/10.1037/bul0000021

Andalman, A. S., Burns, V. M., Lovett-Barron, M., Broxton, M., Poole, B., Yang, S. J., Grosenick, L., Lerner, T. N., Chen, R., Benster, T., Mourrain, P., Levoy, M., Rajan, K., & Deisseroth, K. (2019). Neuronal Dynamics Regulating Brain and Behavioral State Transitions. *Cell*, *177*(4), 970-985.e20. https://doi.org/10.1016/j.cell.2019.02.037

Anderson, D. J. (2016). Circuit modules linking internal states and social behaviour in flies and mice. *Nature Reviews Neuroscience*, *17*(11), 692-704. https://doi.org/10.1038/nrn.2016.125

Anderson, D. J., & Adolphs, R. (2014). A Framework for Studying Emotions across Species. *Cell*, *157*(1), 187-200. https://doi.org/10.1016/j.cell.2014.03.003

Anderson, D. J., & Perona, P. (2014). Toward a Science of Computational Ethology. *Neuron*, *84*(1), 18-31. https://doi.org/10.1016/j.neuron.2014.09.005

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Review of Psychology*, *58*, 373-403. https://doi.org/10.1146/annurev.psych.58.110405.085709

Bateson, M., Desire, S., Gartside, S. E., & Wright, G. A. (2011). Agitated honeybees exhibit pessimistic cognitive biases. *Current Biology: CB*, *21*(12), 1070-1073. https://doi.org/10.1016/j.cub.2011.05.017

Bickle, J. (2016). Revolutions in Neuroscience : Tool Development. *Frontiers in Systems Neuroscience*, *10*, 24. https://doi.org/10.3389/fnsys.2016.00024

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of Animal Consciousness. *Trends in Cognitive Sciences*, *24*(10), 789-801. https://doi.org/10.1016/j.tics.2020.07.007

Boyd, R. (1999). Homeostasis, Species, and Higher Taxa. In R. A. Wilson (Éd.), *Species : New Interdisciplinary Essays* (p. 141-185). MIT Press.

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263-1268. https://doi.org/10.1038/nn1525

Burkhardt, R. W. (2005). *Patterns of behavior : Konrad Lorenz, Niko Tinbergen, and the founding of ethology*. University of Chicago Press.

Calhoun, A. J., & Murthy, M. (2017). Quantifying behavior to solve sensorimotor transformations : Advances from worms and flies. *Current Opinion in Neurobiology*, *46*, 90-98. https://doi.org/10.1016/j.conb.2017.08.006

Calhoun, A. J., Pillow, J. W., & Murthy, M. (2019). Unsupervised identification of the internal states that shape natural behavior. *Nature Neuroscience*, *22*(12), 2040-2049. https://doi.org/10.1038/s41593-019-0533-x

Churchland, P. S. (1986). *Neurophilosophy : Toward a unified science of the mind-brain*. MIT Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575-594. https://doi.org/10.1080/09515080903238930

Craver C (2021) Toward an epistemology of intervention. Optogenetics and maker's knowledge. In J. Bickle, C.F. Craver, A-S. Barwich (Éd.), The tools of neuroscience experiment. Philosophical and scientific perspectives (p. 152-175). Routledge.

Datta, S. R., Anderson, D. J., Branson, K., Perona, P., & Leifer, A. (2019). Computational Neuroethology : A Call to Action. *Neuron*, *104*(1), 11-24. https://doi.org/10.1016/j.neuron.2019.09.038

Deisseroth, K. (2021). *Projections : A story of human emotions* (First edition). Random House.

Deonna, J. and Teroni, F. 2015. Emotions as attitudes. *Dialectica*, 69/3, p. 293-311.

Deutsch, D., Pacheco, D., Encarnacion-Rivera, L., Pereira, T., Fathy, R., Clemens, J., Girardin, C., Calhoun, A., Ireland, E., Burke, A., Dorkenwald, S., McKellar, C., Macrina, T., Lu, R., Lee, K., Kemnitz, N., Ih, D., Castro, M., Halageri, A., … Murthy, M. (2020). The neural basis for a persistent internal state in Drosophila females. *eLife*, *9*, e59502. https://doi.org/10.7554/eLife.59502

Faucher, L., & Forest, D. (Éds.). (2021). *Defining mental disorder : Jerome Wakefield and his critics*. The MIT Press.

Flavell, S. W., Gogolla, N., Lovett-Barron, M., & Zelikowsky, M. (2022). The emergence and influence of internal states. *Neuron*, *110*(16), 2545-2570. https://doi.org/10.1016/j.neuron.2022.04.030

Griffiths, P. E. (1999). Squaring the Circle : Natural Kinds with Historical Essences. In R. A. Wilson (Éd.), *Species : New Interdisciplinary Essays* (p. 209-228). MIT Press.

Griffiths, P. E. (2004). Emotions as Natural and Normative Kinds. *Philosophy of Science*, *71*(5), 901-911. https://doi.org/10.1086/425944

Gründemann, J., Bitterman, Y., Lu, T., Krabbe, S., Grewe, B. F., Schnitzer, M. J., & Lüthi, A. (2019). Amygdala ensembles encode behavioral states. *Science*, *364*(6437), eaav8736. https://doi.org/10.1126/science.aav8736

Hinde, R. A. (1956). Ethological Models and the Concept of « Drive ». *The British Journal for the Philosophy of Science*, *6*(24), 321-331.

Hoopfer, E. D., Jung, Y., Inagaki, H. K., Rubin, G. M., & Anderson, D. J. (2015). P1 interneurons promote a persistent internal state that enhances inter-male aggression in Drosophila. *eLife*, *4*, e11346. https://doi.org/10.7554/eLife.11346

Ji, N., Madan, G. K., Fabre, G. I., Dayan, A., Baker, C. M., Kramer, T. S., Nwabudike, I., & Flavell, S. W. (2021). A neural circuit for flexible control of persistent behavioral states. *eLife*, *10*, e62889. https://doi.org/10.7554/eLife.62889

Jung, Y., Kennedy, A., Chiu, H., Mohammad, F., Claridge-Chang, A., & Anderson, D. J. (2020). Neurons that Function within an Integrator to Promote a Persistent Behavioral State in Drosophila. *Neuron*, *105*(2), 322-333.e5. https://doi.org/10.1016/j.neuron.2019.10.028

Kanwal, J. K., Coddington, E., Frazer, R., Limbania, D., Turner, G., Davila, K. J., Givens, M. A., Williams, V., Datta, S. R., & Wasserman, S. (2021). Internal State : Dynamic, Interconnected Communication Loops Distributed Across Body, Brain, and Time. *Integrative and Comparative Biology*, *61*(3), 867-886. https://doi.org/10.1093/icb/icab101

Kennedy, A., Kunwar, P. S., Li, L., Stagkourakis, S., Wagenaar, D. A., & Anderson, D. J. (2020). Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature*, *586*(7831), 730-734. https://doi.org/10.1038/s41586-020-2728-4

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior : Correcting a Reductionist Bias. *Neuron*, *93*(3), 480-490. https://doi.org/10.1016/j.neuron.2016.12.041

Kuhn, T. S. (1973). Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension : Selected Studies in Scientific Tradition and Change* (p. 320-339). University of Chicago Press.

Liu, M., Kim, D.-W., Zeng, H., & Anderson, D. J. (2022). Make war not love : The neural substrate underlying a state-dependent switch in female social behavior. *Neuron*, *110*(5), 841-856.e6. https://doi.org/10.1016/j.neuron.2021.12.002

Lovett-Barron, M., Andalman, A. S., Allen, W. E., Vesuna, S., Kauvar, I., Burns, V. M., & Deisseroth, K. (2017). Ancestral Circuits for the Coordinated Modulation of Brain State. *Cell*, *171*(6), 1411-1423.e17. https://doi.org/10.1016/j.cell.2017.10.021

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1-25. https://doi.org/10.1086/392759

Marques, J. C., Li, M., Schaak, D., Robson, D. N., & Li, J. M. (2020). Internal state dynamics shape brainwide activity and foraging behaviour. *Nature*, *577*(7789), 239-243. https://doi.org/10.1038/s41586-019-1858-z

Mathis, M. W., & Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, *60*, 1-11. https://doi.org/10.1016/j.conb.2019.10.008

Mill, J. S. (1843). *A system of logic : Ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation* (Repr). Univ. Press of the Pacific.

Murphy, D., & Woolfolk, R. L. (2000). The Harmful Dysfunction Analysis of Mental Disorder. *Philosophy, Psychiatry, & Psychology*, *7*(4), 241-252.

Otchy, T. M., Wolff, S. B. E., Rhee, J. Y., Pehlevan, C., Kawai, R., Kempf, A., Gobes, S. M. H., & Ölveczky, B. P. (2015). Acute off-target effects of neural circuit manipulations. *Nature*, *528*(7582), Article 7582. https://doi.org/10.1038/nature16442

Parker, D. (2018). Kuhnian revolutions in neuroscience : The role of tool development. *Biology & Philosophy*, *33*(3), 17. https://doi.org/10.1007/s10539-018-9628-0

Pessoa, L. (2018). Emotion and the Interactive Brain : Insights From Comparative Neuroanatomy and Complex Systems. *Emotion Review*, *10*(3), 204-216. https://doi.org/10.1177/1754073918765675

Putnam, H. (1967). The nature of mental states. *Art, mind and religion*, 37-48.

Robson, D. N., & Li, J. M. (2022). A dynamical systems view of neuroethology : Uncovering stateful computation in natural behaviors. *Current Opinion in Neurobiology*, *73*, 102517. https://doi.org/10.1016/j.conb.2022.01.002

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178. https://doi.org/10.1037/h0077714

Scarantino, A. (2012). How to Define Emotions Scientifically. *Emotion Review*, *4*(4), 358-368. https://doi.org/10.1177/1754073912445810

Silva, A.J., Landreth, A. & Bickle, J. (2013) Engineering the next revolution in neuroscience. Oxford university press.

Slater, M. H. (2015). Natural Kindness. *The British Journal for the Philosophy of Science*, *66*(2), 375-411. https://doi.org/10.1093/bjps/axt033

Tinbergen, N. (1951). *The study of instinct*. Clarendon Press/Oxford University Press.

Urai, A. E., Doiron, B., Leifer, A. M., & Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, *25*(1), 11-19. https://doi.org/10.1038/s41593-021-00980-9

Wakefield, J. C. (1992). The concept of mental disorder : On the boundary between biological facts and social values. *American Psychologist*, *47*(3), 373-388. https://doi.org/10.1037/0003-066X.47.3.373

Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination. In L. Soler, E. Trizio, T. Nickles, & W. Wimsatt (Éds.), *Characterizing the Robustness of Science* (Vol. 292, p. 61-87). Springer Netherlands. https://doi.org/10.1007/978-94-007-2759-5_2