Interdisciplinary Evolution Research 8

José Manuel Viejo
Mariano Sanjuán *Editors*

Life and Mind

New Directions in the Philosophy of Biology and Cognitive Sciences

Springer

Interdisciplinary Evolution Research

Volume 8

José Manuel Viejo • Mariano Sanjuán
Editors

Life and Mind

New Directions in the Philosophy of Biology and Cognitive Sciences

Springer

Editors
José Manuel Viejo
Department of Logic and Philosophy of Science
Autonomous University of Madrid
Madrid, Spain

Mariano Sanjuán
Department of Logic and Philosophy of Science
Autonomous University of Madrid
Madrid, Spain

ISSN 2199-3068 ISSN 2199-3076 (electronic)
Interdisciplinary Evolution Research
ISBN 978-3-031-30303-6 ISBN 978-3-031-30304-3 (eBook)
https://doi.org/10.1007/978-3-031-30304-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Philosophers have sought to gain insight from science in two distinct ways. Firstly, certain scientific theories have provided valuable perspectives on philosophical questions. For example, to mention one, philosophers studying the nature of space and time have found much to ponder in Einstein's theory of relativity. Secondly, philosophers have also played a key role in shaping scientific research by developing influential theories and initiating key studies. The field of cognitive science offers a prime example of philosophy's impact on science; Fodor's theory of the modularity of mind has had a profound influence on current cognitive architecture theories.

The above examples are far from being exhaustive. The intersection of science and philosophy has led to the emergence of interdisciplinary fields, such as the philosophy of biology and the philosophy of cognitive sciences, which have been incredibly valuable in aiding both scientists and philosophers to gain a deeper understanding of the natural world. In both fields, the philosophy of biology and the philosophy of cognitive sciences, two concepts, respectively, gather the reflections around them: mind and life.

Despite the benefits that academic specialization brings in epistemic terms, one of its well-known drawbacks is the widespread forgetfulness it causes about aspects of the world that we do not have time to inquiry carefully and, viewed in perspective, could be of great utility to our specific focus. This volume aims to bridge that gap with regard to the concepts of mind and life. By working together and sharing cognitive resources, the philosophy of biology and the philosophy of the cognitive sciences not only create new tools to better comprehend previously obscure phenomena but also generate new avenues of research to guide future work.

And just as this book aims to progress on the described path, it also represents the culmination of something that began more than a decade ago, and for which this volume marks the celebration of its tenth edition: The Research Workshop on Philosophy of Biology and Cognitive Sciences (PBCS). The chapters that make up this volume were originally presented at that event.

PBCS is an annual encounter of young scholars that aims at bringing together researchers from diverse disciplinary backgrounds including philosophy, cognitive

science, and biology, to foster discussion and collaboration on shared research interests. The Workshop aims to promote new and innovative perspectives and provide a platform for interdisciplinary dialogue on the intersection of evolution and cognition. It is funded by the "Spanish philosophy network APPLY: New trends in applied philosophy. From theoretical philosophy to new challenges of society (RED2018-102695-T)", the research project "MECABIOSOC: Mechanisms in the sciences, from the biological to the social FFI2017-89639-P (2018-2021)" and was elected by the International Society for the History, Philosophy, and Social Studies of Biology (ISHPSSB) as one of their Off-Year Workshops. As mentioned before, this volume is indeed motivated by its tenth edition, hosted by the Autonomous University of Madrid (UAM) on Zoom on May 3rd–4th, 2021.

Throughout its ten-year history, the PBCS has attracted over a hundred of contributed talks, as well as keynote speakers such as Cristian Saborido, Laura Nuño de la Rosa, David Ward, Xabier Barandiaran, Glenda Satne, Francesca Merlin, Thomas Raleigh, Lynn Chiu, Marc Artiga, Susana Monsó, Cristopher J. Austin, Manuel Heras Escribano, Gaëlle Pontarotti, Javier González de Prado, and Marta Jorbá. They are all now leading researchers in their respective fields and many of them are still attached to the PBCS as either part of the scientific committee or contributors to this volume (as Javier González de Prado and Cristian Saborido).

Since its inception, the PBCS meetings have expanded in both scope and complexity. Initially a small-scale project, it has now become a prominent forum in the field. Every year, the PBCS receives many proposals and draws in hundreds of members from the philosophy of science community, who would likely be interested in this volume. This collected volume represents a significant achievement for the project and its tangible consolidation, and will likely inspire new researchers to become involved in the coming years.

Many people have been directly or indirectly involved in this project. We feel indebted to all of them. Firstly, we want to express our gratitude to Cristian Saborido, unmoved mover of the PBCS, for trusting us with the organization of the congress. Many thanks also to the scientific committee for their involvement in the Workshop, which we know goes beyond academia: Marc Artiga, Leonardo Bich, María Cerezo, Manuel Heras-Escribano, Susana Monsó, Laura Nuño de la Rosa, Manuel de Pinedo, Javier Suárez, and Josefa Toribio.

We would like to extend our sincerest gratitude to the team at Springer's *Interdisciplinary Evolution Research* series. Special mention goes to Nathalie Gontier for her determined trust in our project and Sabine Schwarz and Srinivasan Manavalan for their invaluable assistance, patience, and support throughout the editorial process. Your contributions have been fundamental in the success of this volume.

Thank you also to the new committee of the PBCS who last year took charge of the project: Ana Cuevas Badallo, Mariano Martín-Villuendas, Juan Gefaell, Esther Palacios Mateos, and María del Pilar López. We know that the PBCS is in the best hands.

For this last acknowledgment, I, Mariano Sanjuán, take the pen. I am deeply indebted to José Manuel for the unwavering determination and skill with which he

guided this ship through tumultuous waters. Despite the challenges posed by a global pandemic and numerous other obstacles that we found along the way, his leadership and resilience were key in ensuring the success of this endeavor. It is individuals like him who drive the advancement of knowledge forward.

This volume is not intended as an introduction to the field, but it is nonetheless aimed at a broad audience. Students and researchers in philosophy with a particular interest in life, cognition, and evolution, as well as biologists and cognitive scientists, should find the chapters of this volume relevant. Additionally, in line with the spirit of this series, this book does not conform to a single academic field. The book is sure to appeal to any reader with an interest in the topics covered.

Madrid, Spain

José Manuel Viejo
Mariano Sanjuán

Contents

Contributors

Arantza Etxeberria Agiriano Department of Logic and Philosophy of Science, University of the Basque Country, Donostia – San Sebastián, Spain

Giuseppe Flavio Artese Institute of Philosophy, University of Kassel, Kassel, Germany

Francesca Bellazzi Department of Philosophy, University of Bristol, Bristol, UK

Mark-Oliver Casper Institute of Philosophy, University of Kassel, Kassel, Germany

David Cortés-García Department of Philosophy, University of the Basque Country, Leioa, Spain

Antonio Danese University of Padua, Padua, Italy

Javier González de Prado Department of Logic, History and Philosophy of Science, National Distance Education University, Madrid, Spain

Juan Gefaell Centro de Investigación Mariña, Departamento de Bioquímica, Genética e Inmunología, Universidade de Vigo, Vigo, Spain

Malena León Institute of Humanities, CONICET-National University of Córdoba, Córdoba, Argentina

Ignacio Maeso Andalusian Centre for Developmental Biology (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain

Tiago Rama Autonomous University of Barcelona, Barcelona, Spain

David Ricote Andalusian Centre for Developmental Biology (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain

Ainhoa Rodriguez-Muguruza Department of Logic and Philosophy of Science, University of the Basque Country, Donostia – San Sebastián, Spain

Tiina Rosenqvist Department of Philosophy, University of Pennsylvania, Philadelphia, PA, USA

Cristian Saborido Department of Logic, History and Philosophy of Science, National Distance Education University, Madrid, Spain

Mariano Sanjuán Department of Logic and Philosophy of Science, Autonomous University of Madrid, Madrid, Spain

José Manuel Viejo Department of Logic and Philosophy of Science, Autonomous University of Madrid, Madrid, Spain

Peter Woodford London, UK

Giacomo Zanotti Politecnico di Milano, Milano, Italy

Life and Mind: An Introduction

Mariano Sanjuán and José Manuel Viejo

Abstract The spectrum that bridges the distance between the philosophy of biology and the philosophy of cognitive sciences is paved with shared concepts such as cognition, evolution, genetics and teleology, among many others. The existence of such shared areas of interest indicates the suitability and significance of volumes like the one presented here. But while the abovementioned concepts are becoming increasingly prominent in researchers' agendas, there is a lack of a cohesive view that brings together these topics in a single compilation. It is for this reason that *Life and Mind* aims to provide a shared platform where both of these disciplines are intertwined and related, fostering new perspectives for thinking philosophically about life and its various aspects in a comprehensive manner.

Keywords Cognition · Evolution · Genetics · Teleology · Philosophy of biology · Philosophy of cognitive sciences

1 Life and Mind: An introduction

The relationship between life and cognition is a multifaceted and constantly evolving subject. However, it can be accepted without controversy that cognition is a feature that emerged in the course of evolution to help organisms survive and reproduce, and specifically, to deal with complex environments, enabling such organisms to navigate and adapt to them (Godfrey-Smith 1998). The evolution of cognition has enabled living entities to perceive, learn (Ginsburg and Jablonka 2019), evolve, communicate and respond to their surroundings. This strong correlation between life and the origins of cognition is undeniable (Barbaras 1999; Margulis 2001; Veit 2022). Though the precise nature of the type of imbrication they maintain remains a point of debate, it is clear that the mind and life are intimately interconnected and essential to our understanding of the living world.

M. Sanjuán (✉) · J. M. Viejo
Department of Logic and Philosophy of Science, Autonomous University of Madrid, Madrid, Spain
e-mail: msanjuans01@larioja.edu.es; jose.viejo@educa.madrid.org

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_1

The authors contributing to this collective work share a similar perspective to the one just described. This is perhaps the central leitmotif of the present volume, one which provides it with a certain structure and coherence: the belief that it is possible to cease relegating the concepts of life and mind to separate sub-specializations and to begin to consider them together, as integral parts of a comprehensive perspective. In particular, four ideas or concepts will serve as examples of this joint vision: human and animal cognition, evolution, genetics, and teleology. These issues are recurrently discussed throughout the volume, with each of the papers that address them offering a distinct viewpoint. We consider this to be a strength of the book: the diversity of its chapters reflects the diversity of current research perspectives on these topics.

Life and Mind is divided into four parts. Each part revolves around a central concept that serves as a common point of reference, and each chapter sheds light on a different aspect of this concept. This gives the book versatility, allowing it to be read in different ways. We would suggest following the order of presentation, starting with the first chapter and concluding with the last. However, the book can also be read selectively, leaving the reader free to jump to specific parts that align with their interests. In what follows, we summarize each of the chapters that will appear next.

2 Introduction to Part I: Embodiment, Perception and Cognition

In his analysis of the ongoing debate on animal understanding, **Peter Woodford** aims to shed light on the type of understanding that holds significance in discussions surrounding self-awareness. Woodford argues in his article **'Animal Understanding and Animal Self-Awareness'** that by examining the nature of animal understanding, it is possible to identify a justifiable concept, which makes the idea of non-human self-understanding more viable. Woodford's argumentation proceeds in a number of steps. First, he analyzes the nature of animal understanding by drawing on recent discussions in epistemology. These discussions define understanding in terms of a grasp of causal patterns that hold in the world. Using examples from research on animal cognition and behavior, he defends a concept of animal self-understanding as a grasp of causal patterns that hold between an animal's own actions and the effects of its actions in the world. He then shows that recently developed concepts of animal bodily and social self-awareness, which have highlighted the importance of various forms of agency in the concept of animal self-awareness, can inform and further develop the plausibility of animal self-understanding. The chapter concludes by showing how the concept of self-understanding defended by Woodford fits into recently articulated evolutionary views of the emergence of sentience and self-awareness.

Woodford's chapter makes clear that cognition cannot be understood as a mere intracranial form of information processing; rather, the environment and the agent's

actions within it must also be taken into account. Setting aside distinctions between animal and human self-awareness, the second chapter delves further into enactivism, where **Mark-Oliver Casper** and **Giuseppe Flavio Artese** find an attractive alternative to mainstream paradigms in the cognitive sciences. In their chapter '**A Methodological Response to the Motley Crew Argument. Explaining Cognitive Phenomena through Enactivism and Ethology',** they claim that ethological investigations provide a solid methodical grounding for enactivism, and offer important insights into the phenomena that situated cognition researchers would not hesitate to define as cognitive. Casper and Artese provide a solid defense against what they regard as one of the main objections to the enactive theory of cognition: the 'Motley-Crew Argument', the idea that the vast number of entities that enactivism identifies as relevant for studying cognitive phenomena seem to be ultimately unamenable to rigorous scientific scrutiny. In their chapter, they maintain that such challenges can be addressed by the enactive approach to cognition, by paving the way for implementing a suitable methodology that comes from another but theoretically adjacent field, i.e. the methods applied in the field of biological ethology. As they noticeably show by appealing to the concepts of action-readiness and dynamically emerging interaction patterns, as well as case studies such as flight initiation distance in gregarious birds, enactivism and ethology have strong commonalities and very often overlap in their usage of concepts. If this is accepted, Casper and Artese claim, a path opens up for enactivists to answer the Motley-Crew argument.

In addition to enactivism, another topic of long-term and continued relevance is the nature of mental properties. Working on empirical evidence concerning the electrochemical mechanisms that underlie the workings of neurons, **Giacomo Zanotti** takes the baton with his chapter **'Causal Closure, Synaptic Transmission, and Emergent Mental Properties'**, in order to threaten the physicalist's reliance on the so-called Causal Closure principle (i.e. the idea that if a physical event has a sufficient cause, then it has an immediate sufficient physical cause). Even though the more neuroscientists and psychologists succeed in uncovering physical causes that might give rise to mental properties, the less room is left for the exercise of the causal powers *of sui generis* mental properties, Zanotti claims that physicalists still need to assume that the behavior of the nervous system is compositionally determined by the behavior of its working units. In other words, the physicalist has to presuppose that the causal power of the nervous system is nothing but the sum of the causal powers of its organized components, which are neurotransmission mechanisms; and justifying this assumption, according to Zanotti, is not possible at present.

Whether they are physical, epiphenomenal or otherwise, we all have color experiences. In the fourth chapter, **Tiina Rosenqvist** keeps coloring our volume with '**Color and Competence: A New View of Color Perception'.** Here she sketches a new view of color perception that centers on the notion that color vision is 'competence-embedded'. Rosenqvist's novel position entails two main claims. First, she believes that the overarching function of color vision is to enable and enhance the manifestation of relevant species-specific competences (these being competences, in the case of humans, to enable and enhance figure-ground segregation, object identification and re-identification, and property identification).

Secondly, Rosenqvist's view also implies that color experiences are correct when they result from processing that directly and non-accidentally subserves the manifestation of such competences. With a keen eye, Rosenqvist applies her perspective in order to demonstrate its ability to incorporate and explain a broad range of color perception phenomena, including many challenging cases. She differentiates between ideal and non-ideal cases of color perception. In ideal cases (such as seeing a ripe tomato as red), the demands imposed by the relevant competences line up, and the color visual system can simultaneously fulfill its enhancement function concerning all of them. In non-ideal cases (such as color illusions), in contrast, the demands of the relevant competences diverge and clash. Using Akiyoshi's optical illusions as an example, Rosenqvist shows that the apparent strangeness of such cases is a consequence of stimulus properties that pit the demands of the competences against one another. In the last part of her chapter, she convincingly argues that her view not only comes with a lot of explanatory purchase, but also allows us to uphold the powerful intuition that color visual systems are generally well-functioning systems that nevertheless sometimes fail.

Bringing back the 4E cognition framework, **Ainhoa Rodriguez-Muguruza** explores applying the idea of desynchronization to disrupted interactions between menstruating bodies and their environment, and reclaiming its adequate understanding as a way towards re-synchronization, in **'Menstrual Cycles as Key to Embodied Synchronisation'.** Here she presents menstrual rhythms as an essential constitutive aspect of the embodied experience of menstruating female bodies, crucial for philosophical explorations of embodiment and of its consequences for somatic and mental health. She aims to underline its infradian nature and to analyze whether failing to acknowledge this biological rhythm could trigger further pathologies. Here she deems phenomenological accounts of cognition and the paradigm of 4E cognition to be crucial for discussions about intersubjectivity and embodied interaction, and in particular for showing how social and environmental factors can disrupt the harmonious interactions behind menstrual functions, leading to the individual's desynchronization not only with their reproductive clock, but also with other organismic processes. By the end of the chapter, Rodriguez-Muguruza also explores the potential link between these disruptions and mental pathologies whose prevalence has increased within women with menstrual disorders.

3 Introduction to Part II: Evolution, Language and Culture

The second part of the volume, dedicated to a wide range of topics related to evolution, opens with '**Is Cultural Selection Creative?'** by **Malena León,** where an innovative way of accounting for creative processes that incorporates factors beyond the individual's cognitive abilities (with a special emphasis on evolutionary-cultural processes) is proposed. Creativity, traditionally regarded as a subproduct of a brilliant mind, is in fact the result of the 'collaboration' of innumerable people who may not even know each other or completely understand the processes to which they

contribute. This conceded, there is a need to delve further into the implications that arise from theories of cultural evolution for a theoretical approach to creativity, so as to strengthen the explanatory links between theories of cultural evolution and theories of creativity. For that purpose, León draws on a dispute that has taken place in evolutionary biology and the philosophy of biology: the debate on whether natural selection is a creative force. To argue that cultural selection can be creative, she analyzes the arguments that have been used in the literature, extrapolating into the field of culture those criteria that are used in evolutionary biology. Finally, she tests these criteria by analyzing real-world examples to show that cultural selection behaves, at least sometimes, in a creative way.

Just as León's chapter illustrates how case studies from science can offer compelling empirical evidence to bolster philosophical discourses, **Juan Gefaell** and **Cristian Saborido** demonstrate how philosophy can help us unravel and better understand the entangled conceptual connections between different scientific theories. In this vein, **'Incommensurability in Evolutionary Biology: The Extended Evolutionary Synthesis Controversy'** revisits and criticizes Pigliucci's (2017) analysis of the relation between the Modern Synthesis (MS) and the Extended Evolutionary Synthesis (EES). Taking up Kuhn's well-known concept of incommensurability, Pigliucci argues that the MS and the EES are in fact commensurable frameworks (at the methodological, observational and semantical levels) and that their relationship is best understood in terms of the EES being a business-as-usual extension of the MS. In contrast, Gefaell and Saborido hold that Pigliucci's analysis of incommensurability is limited because he seems to assume that incommensurability is a holistic phenomenon, he overlooks the most contentious issues in the controversy between the MS and the EES, he does not provide a sound interpretation of observational incommensurability, and he contends that incommensurability always implies a paradigm shift. After highlighting Pigliucci's difficulties, Gefaell and Saborido provide an alternative interpretation of incommensurability between the MS and the EES that seeks to overcome Pigliucci's limitations. As they persuasively argue, the MS and the EES are in fact methodologically, observationally and semantically incommensurable. By the end of the chapter, Gefaell and Saborido discuss which mode of scientific change better explains the current situation in evolutionary biology, arguing that it is too soon to make any definite statement about it, but leaving the door open to alternative approaches to several philosophical problems related to the rationality of scientific change, scientific realism, and how to deal with deep disagreements within a given scientific community.

The Modern Evolutionary Synthesis and the more recent Extended Evolutionary Synthesis remain as protagonists in the next chapter of the volume. In **'Ontologies in Evolutionary Biology: The Role of the Organism in the Two Syntheses', David Cortés-García** and **Arantza Etxeberria Agiriano** examine how organisms and their role in biological phenomena have varied immensely from the onset of Darwinian thinking to the development of the aforementioned theories. Here they argue that whereas the Modern Synthesis became increasingly reductionist and monist, the Extended Synthesis is constituted by a varied array of models that are able to accommodate different ontological levels, whereby the organism stands as crucial

at the crossroads of many other significant ontological aspects, because of its flexibility and potential inclusiveness. In their meticulous examination of the evolution of the concept of an organism in the context of post-Darwinian evolutionary biology, Cortés-García and Etxebarria Agiriano conclude that the reliance on adaptation-based explanations, which assumed that organisms are nothing but a collection of adaptations, ended up excluding the organism from being considered a worthwhile notion in evolutionary biology. Yet, they claim, failure to reduce the systemic nature and ecological dynamics of the organism (including its properties of agency and organization) to the framework of the Modern Synthesis imposes some important drawbacks. The authors then examine two of the main ontological objections to the Modern Synthesis framework, in order to conclude that a strong, organizational, relational, and agential notion of an organism becomes inevitable for understanding many phenomena without which evolutionary biology is incomplete. Finally, in relation to the historical relations within evolutionary biology, Cortés-García and Etxebarria Agiriano support the idea that rather than the successive unifications or expansions of the theoretical framework that are usually presumed, it is the scientific activity of the actors involved in evolutionary explanations that displayed a varied set of research questions and gave rise to the network-like array of models and practices that eventually constituted evolutionary biology, whose epistemological aspects are importantly influenced by the ontologies that different theories may commit to.

4 Introduction to Part III: Gene and Genotype Metaphysics

The next two chapters of the volume shed new light on different philosophical considerations about genetics. Specifically, these chapters delve into the discussion of the types of entities that are referred to when concepts such as genotype, phenotype or gene are used. To begin with, **David Ricote** and **Ignacio Maeso** reformulate the meaning of 'genotype' and 'phenotype' in the framework of the type/token distinction, in their chapter entitled **'A New Perspective on Type-Token Distinction in The Genotype and Phenotype Concepts'.** Their view is original and innovative, as they evaluate the type-token relations of both concepts, proposing a new and solid conceptual background in order to differentiate genotype and phenotype by how they classify their tokens. According to their perspective, genotypes should be defined independently of genes, thus distinguishing genotype from *genetype*: genotypes should classify whole inherited structures while a genetype is a way of classifying genes. After critically reviewing the main definitions of gene and showing that they are not suitable for defining the tokens of a genotype, Ricote and Maeso suggestively propose that instead of genes, the material instances classified by genotypes—the *genotokens*—should be complete structures that are inherited in each cell reproductive cycle, the most straightforward example being the whole DNA or genome sequence of a cell. These molecular structures can be clearly distinguished as units, in clear contrast to genes, which cannot be easily

delimited as units, or equivalently as tokens. Once this has been done, Ricote and Maeso propose that genotypes could be considered to be natural kinds, because they reproduce and conserve their type-identity by self-templated replication; while in contrast, they argue, phenotypes classify tokens (*phenotokens*), depending on the intentional individuation and classifications of humans; and therefore, they argue, phenotypes do not constitute natural kinds. Precisely this topic (i.e. natural kinds in the philosophy of genetics) is discussed extensively in the subsequent chapter.

There is no denying that much ink has been spilled discussing the nature of our concepts and, in particular, those used in science: whether they describe actual facts of the world or are simply tools of conceptual clarification. The development of science provides us with new objects for reconsidering these classic questions. This is exactly what **Francesca Bellazzi** does in the following chapter, focusing on the concept of a molecular gene. In **'The Gene as a Natural Kind',** she applies Khalidi's (2013) definition of natural kinds as projectible categories and nodes in causal networks, and takes a stance on this matter, siding with the realists. In her approach, the category of 'molecular gene' used in scientific practice corresponds to a natural kind, despite the complexity of the properties that characterize it, and it captures some objective features of reality. Bellazzi's chapter has important implications. First, a better understanding and comprehension of whether something is a natural kind or not is essential, because the naturalness of a given category can provide us with a further justification for why we can make more robust inferences from it. In doing so, identifying something as a natural kind can support the justification of a theory that presents such a kind. Second, a natural kind is more than a theoretical entity whose properties are postulated for practical purposes, and this can direct research into discovering (rather than merely postulating) its features. This supports the role that they also have in the process of discovering new information about such a category. A natural kind corresponds to something objective in the world, meaning that some properties could be discovered as belonging to it, and some could not.

5 Introduction to Part IV: Teleology in Biology and Cognitive Sciences

The last subject of dispute that runs through this volume brings us to a variety of philosophical concerns surrounding the relationship between teleology and natural selection. In the first of two chapters, '**Teleological Explanations and Selective Mechanisms. Biological Teleology Beyond Natural Selection**', **Javier González de Prado** and **Cristian Saborido** vindicate biological teleology from a pioneering point of view. Starting with a general definition of selection as differential reinforcement, they interpret the different types of teleological explanation, both biological and non-biological, as specific cases of selective explanations, of which evolutionary

explanations would be only a specific subset (rather than the only one). They propose, then, to take selection to be more generally a matter of differential reinforcement, in order to claim that selection involves the differential reinforcement of certain effects or traits, where this reinforcement may be a matter of being promoted, reproduced, preserved, stimulated, or intensified somehow. The notion of reinforcement that González de Prado and Saborido propose is sufficiently flexible to cover the great variety of cases of selection, including its most paradigmatic forms. It makes sense to consider natural selection to be a selective process precisely because it involves differential reinforcement, in the form of differential reproductive rates. In light of these considerations, González de Prado and Saborido go on to define a selective mechanism as a mechanism by which the behavior of a system and its relationship with its environment are modified in such a way as to reinforce the presence of certain effects or traits over other alternatives. Finally, they argue that an explanation is teleological if it appeals to the effects of a trait which explain its reinforcement through a selective process—a trait can be teleologically explained if it is structured as the result of a *selective process*—and apply their view to biological regulation as an example. All things considered, Gónzalez de Prado and Saborido's chapter persuasively argues that biological regulation should be considered a selective process, giving rise to its own form of biological teleology.

The volume closes with '**Evolutionary Causation and Teleosemantics**', by **Tiago Rama**, in which he takes up recent disputes about different interpretations of the causal structure of natural selection, in order to conclude that adopting an alternative to mainstream etiological views (i.e. non-causal, statisticalist), opens the door for two lines of thought of special relevance to teleosemantics, and sets out the biological foundations of a new teleosemantic account, which he labels as *Agential Teleosemantics*. As for the first line, Rama claims, a statisticalist reading of natural selection allows for setting different challenges for etiological teleosemantics. In this sense, rejecting the causalist reading of natural selection is tied to different challenges to the foundations of modern evolutionary theory. The second implication Rama highlights in his chapter is that his alternative approach suggests an individual-level view of the causes of evolution. According to this view, all causes that produce apt and complex living systems are individual causes. This view of natural selection is connected to different contemporary approaches in the philosophy of biology that stress the central explanatory role of organisms in biological theory.

6 Conclusion

In this introduction, the main theses and arguments of each chapter of this compilation have been presented. Now readers are in a position to verify for themselves, as warned before, the plurality of themes and approaches that it comprises. This fact has two readings. First, this volume clarifies the richness in methods, approaches, themes and issues that philosophy of biology and philosophy of cognitive sciences enjoy.

This bodes well for future research in these areas. Secondly, the idiosyncrasy of this compilation can be seen as a regulatory ideal for how to continue building knowledge in related fields. We consider, as mentioned at the start, that it is beneficial to our understanding of the natural world to share resources and spaces. It is also worth noting again that the chapters are self-contained and can therefore be read without following their numerical order. We hope that the reader enjoys each of them and returns, whenever curiosity or the desire to know demands it, to *Life and Mind*.

References

Barbaras R (1999) The movement of the living as the originary foundation of perceptual intentionality. In: Petitot J, Varela FJ, Pachoud B, Roy JM (eds) Naturalizing phenomenology: issues in contemporary phenomenology and cognitive science. Stanford University Press, Stanford, CA, pp 525–538

Ginsburg S, Jablonka E (2019) The evolution of the sensitive soul. MIT Press, Cambridge, MA

Godfrey-Smith P (1998) Complexity and the function of mind in nature. Cambridge University Press, Cambridge

Khalidi AM (2013) Natural kinds and human categories. Cambridge University Press, Cambridge

Margulis L (2001) The conscious cell. In: Marijuán PC (ed) Cajal and consciousness: scientific approaches to consciousness on the centennial of Ramon y Cajal's Textura, Annals of the New York Academy of Sciences, vol 929. New York Academy of Sciences, New York, pp 55–70

Pigliucci M (2017) Darwinism after the modern synthesis. In: Delisle RG (ed) The Darwinian tradition in context: research programs in evolutionary biology. Springer International Publishing, Cham, pp 89–103

Veit W (2022) Complexity and the evolution of consciousness. Biol Theory. https://doi.org/10.1007/s13752-022-00407-z

Part I
Embodiment, Perception and Cognition

Animal Understanding and Animal Self-Awareness

Peter Woodford

Abstract Theorists arguing that non-human animals (simply animals from this point forward) are self-aware often make the case on the basis that non-human species *understand* aspects of themselves and the world, and these forms of understanding indicate self-awareness. But the notion of understanding in this context is often taken for granted. This article aims to analyse the nature of animal understanding to clarify the kind of understanding that matters for discussions of self-awareness, namely, self-understanding. I proceed by drawing on discussions of understanding offered in contemporary epistemology, and then by discussing the relevance of the concept of self-understanding here for discussions of animal self-awareness. I argue that the kind of self-understanding relevant to discussions of animal self-awareness is specifically an animal's understanding of its own causal influence on the world and on others.

Keywords Animal minds · Animal cognition · Comparative psychology · Philosophy of biology · Self-awareness

1 Introduction

While recent work in philosophy and the sciences has defended the notion that non-human animals and pre-verbal infants understand the world, the idea that animals understand themselves seems to appear less plausible (Grimm 2016; Baumberger et al. 2017). The aim of this article is to show that there is a defensible concept of self-understanding that allows for the recognition of non-human forms of self-understanding. I do not aim to go all the way to showing that non-human animals *do* understand themselves, but rather to articulate a concept of self-understanding that opens space for the possibility of non-human self-understanding. Moreover, I argue that this is the form of understanding that is relevant to the case for

P. Woodford (✉)
London, UK

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_2

animal self-awareness. The article proceeds in a number of steps. First, I analyse the nature of animal understanding by drawing on recent discussions in epistemology. Next, I defend a concept of animal self-understanding that is based on these discussions of animal understanding. I then show that recent discussions of animal self-awareness, which have highlighted the importance of various forms of agency – particularly bodily and social – in the concept of animal self-awareness, can inform and further develop the concept of animal self-understanding. Finally, I show how the concept of self-understanding defended here fits into recently defended evolutionary views of the emergence of sentience and self-awareness.

2 Understanding and Animal Understanding

What is understanding? The philosophical literature on understanding is vast, and the landscape of debate shows disagreement over fundamental conceptual questions. Nonetheless, some consensus seems evident around the notion that understanding is something more than knowledge, if knowledge consists in the justified affirmation of true beliefs. A prevalent conception of understanding is that it is a grasp of causal relations and dependencies that hold within various arenas of the world (Zagzebski 2001). To take an example borrowed from a recent article, one might know *that* a house caught on fire due to faulty wiring, without *understanding* how a house might catch on fire due to faulty wiring (Grimm 2016). Thus, unlike justified affirmation of propositions that are true, understanding involves knowledge of how things *work*; someone who understands can grasp the patterns and causal dependencies in the world that, at various levels, undergird the facts we might affirm. Another common example concerns the familiar experience of taking a test. Some students might answer questions correctly on an exam by rote memorisation of a list of true facts, affirmed through justified trust in the epistemic testimony of educators, without really *understanding* the reasons why these facts obtain. So, I can pass the test in biology without really understanding much about the way the biological world works. According to this line of thought, understanding can involve propositional knowledge, but it is better thought of as a grasp of how things *work* in a given region of world.

Given this account, it might appear that the case for animal understanding is not very strong. There are two reasons this might be the case: first, understanding causal patterns is often thought of as *scientific* understanding, and, second, such understanding still requires propositional knowledge 'that' certain facts obtain. Let us consider these in turn. It might appear that true understanding of causal relations is involved in *scientific* understanding, or the kind of knowledge of nature's causal structure that has been extremely hard won through the advancement of the natural sciences. To *really* understand the event of a house catching on fire, we need to know something about physics, chemistry, and thermodynamics. Or, to understand why a lake is frozen, we need to know about geological patterns, meteorology, and the earth's orbit. If such sophisticated forms of culturally preserved and transmitted

causal understanding are required to *really* understand, then the case for understanding in animals appears a non-starter. However, some epistemologists have recently pushed against an account of understanding that sets the bar at such fine-grained, extensive knowledge of causal relations (Grimm 2016; Baumberger et al. 2017). For example, if a plastic bag is rolling along the ground and flying up in the air, I understand that it is the strong wind that is moving it. If a lake is frozen, I understand that it is due to the cold. These explanations may be a coarser, less fine-grained understanding of causal patterns than many today have come to expect, but it is the basic understanding that sciences then go on to deepen and clarify. The point here is that my understanding need not extend as far as the causes of the wind itself, or of the cold, for me to properly understand what is going on with the lake or the plastic bag. Still, both mechanistic, fine-grained causal understanding of nature in the sciences and more 'everyday' understanding of how the world works suggest that my understanding is a matter of grasping causal patterns and relations of dependency among phenomena. As some epistemologists have argued recently, understanding is a matter of grasping how some region of the world works (Grimm 2016).

The second aspect of the problem might appear more difficult. If understanding of causal relations requires mastery of the concept of *causation*, or, further, of propositions asserting various states of affairs and causal dependencies between them, then it might also appear unlikely that animals possess understanding. Here we encounter something of a Hume versus Kant problem. Can an understanding of causal relations and dependencies be indicated by the mere lack of surprise that one event has followed another? Can it be indicated by the mere confident *expectation* that some event will occur after another? Expectations can, of course, involve concepts, but does understanding require grasp of concepts, for example the concept of one event making another event happen? There is not space to settle this debate here, but let me offer something of a compromise to get the rest of the inquiry going. We can grant that there is certainly a difference between 'Kantian' conceptual understanding that involves the concept of causation and the understanding indicated by 'Humean' expectation; nevertheless, 'Kantian' understanding involving the concepts need not be thought as the measure of understanding *tout court*.

It is certainly difficult to accept the notion that animals understand causal patterns in the world at the level of the natural sciences. But there are well known examples of animal behaviour that support the ascription of such 'Humean' understanding to animals. Experimental research on rats (Blaisdell et al. 2006), corvids (Taylor et al. 2012; Jelbert et al. 2019), primates (Völter et al. 2016), and on pre-verbal children (Gopnik et al. 2001; Kushnir and Gopnik 2005) suggests that non-verbal animals do indeed track accurately the causal relations and patterns between events. To take a recent representative example, Jelbert et al. 2019 tested whether New Caledonian crows could infer whether an object was light or heavy based on whether it could be moved by a breeze generated by a fan. Crows were trained to receive a reward after dropping either a light or a heavy object into a dispenser. Then, birds were able to watch how two suspended objects behaved in front of a fan. They were able to pick out the light or heavy object accurately (whichever they were trained to expect a reward from) 73% of the time, and did no better than chance in control trials without

the fan on. Researchers argue that the crows were rightly inferring the weight of objects by watching how they behaved in front of a fan (Jelbert et al. 2019). Other experiments with crows suggest that they can also infer *agential* causes behind the observation of some phenomena, which has also already been shown in pre-verbal children (Saxe et al. 2005; Saxe et al. 2007; Taylor et al. 2012).

Such experimental results suggest a new possibility. We can recognise that scientific understanding is more extensive and detailed than 'everyday' understanding, and that humans may possess a form of Kantian conceptual understanding that Humean animals lack. Nonetheless, 'everyday' Humean understanding can be recognised *as* understanding of causal relations, dependencies, and patterns, and this gives us what we need to make the case for animal understanding (Grimm 2016). Some might argue that experiments like the one cited above show that animals do possess non-linguistic concepts as well – even if the concept of causation is not necessarily one of these concepts. This is a debate worth having, but my focus at this stage of the argument is only to defend the plausibility of a conception of understanding that involves only the ability to track causal patterns accurately and to develop expectations about them.

3 Understanding and Self-Understanding

If the development of expectations about patterns in the world that do, indeed, match such patterns counts as a basic form of understanding, we can also suggest some implications for the notion of self-understanding. Divergences in the concept of self-understanding mirror those of the concept of understanding. For example: do I need to have scientific understanding of causal patterns in my body, or about how I will perform in psychological experiments to have self-understanding? Just as in the case of a frozen lake, we can distinguish between more or less fine-grained knowledge here. Let us say I do not like cilantro, and I do not like it because it leaves a soapy taste in my mouth. I understand that I have no desire for cilantro because every time I taste it, it leaves a soapy taste in my mouth. I do not need to understand biochemistry or the physiological mechanisms of taste sensation, or even that a gene has been isolated that appears to explain variation in the taste of cilantro (Eriksson et al. 2012). I understand that it tastes bad, and so I avoid it; this too constitutes a form of 'everyday' understanding of causal patterns involving the traffic between my sensations and the world. Of course, we can admit that there are more or less sophisticated forms of self-understanding, but such an 'everyday' understanding of how the world occasions my own sensations is understanding nonetheless.

Do animals have this 'everyday' sort of self-understanding? To stick with the example above, the difficulty here is to determine whether a behavioural response, such as avoiding eating something that an animal has tasted, involves an awareness of a causal pattern. Experimental research on the phenomenon of *conditioned taste avoidance* (CTA) suggests that avoiding food that has been manipulated to taste bad or toxic is widespread across a variety species including mantises, blue jays, slugs,

and molluscs (Bures et al. 2002; du Toit et al. 1991; Parker et al. 2008; Reilly and Schachtman 2008). The survival value of striking an optimal balance between openness to novel foods and avoidance of toxic foods is thought to explain the prevalence of this phenomenon and its primitive neurological underpinnings. Sensations that are experienced after eating a food lead a creature to develop expectations that steer them away from foods in the future. Examples of CTA can involve classical conditioning in which an animal is trained to expect a desired food or to exhibit vomiting in the expectation of a toxic food in response to an arbitrary stimulus (such as ringing a bell); however, CTA is often studied as a more particular phenomenon in which an internal state of the organism – such as pleasure, equilibrium, or disequilibrium – serves as either the reward or punishment for ingesting a food source (Bures et al. 2002).

Studies of CTA, I think, suggest that animals can develop an understanding of how *they* work through an understanding of causal patterns involving their own sensations. Just as I arrive at a form of self-understanding by tasting cilantro, so might animals come to understand the traffic between their bodily responses and the world. We can ask about the relation between awareness and understanding for such cases as well. Here, it also appears that awareness is a necessary, but not sufficient, condition for self-understanding; furthermore, understanding is a sufficient, but not necessary, indicator of awareness. If a creature is not aware of its own bodily states, psychological states, or behaviours, it cannot be said to understand causal patterns that involve them.

Just as animals need to be aware of some features of the external world to understand causal patterns therein, so too do they need to be sentient and aware of the stuff in the world that occasions their own sensations. These need not be complex sensations; it seems that basic experiences of pain and pleasure are enough. Philosophers working on the nature of animal minds have argued that the most basic form of self-awareness is *bodily self-awareness* (DeGrazia 2009; Bermudez 1998). While bodily self-awareness is thought to involve more than just sentience (more on that in a moment), the ability to experience basic sensations such as pleasure or pain is required. For example, David DeGrazia argues that bodily self-awareness can involve anything from temporally immediate experiences of pain, pleasure, thirst, hunger to more complex emotions such as fear in anticipation of danger, or excitement in anticipation of something pleasurable. These more complex forms of awareness involve memory and/or projection, and some examples of conditioned taste avoidance may belong in this category insofar as they involve anticipation or expectation. The main claim here is that self-understanding ought to be understood as the ability to track accurately causal relations involving one's own sensations and the world.

4 Self-Understanding and Agency

The previous sections made the case that animals should be said to understand themselves if they develop accurate expectations about causal patterns involving the traffic between the world and their own bodily sensations. Yet, one might worry that only the presence of immediate sensations and the avoidance of noxious stimuli is not enough to constitute self-understanding. After all, such avoidance behaviour might be a simple, automatic response to an environmental stimulus, no different from a thermostat. Examples of conditioned taste avoidance show that organisms can learn to act appropriately – adaptively – as a result of their own sensations. But, given the evolutionary primitiveness and prevalence of this phenomenon across species, including molluscs, something more seems necessary. In addition to having rudimentary forms of subjective experience such as pleasure and pain, as in examples of taste, self-understanding seems to involve the capacity to behave flexibly in response to learning how phenomena 'out there' both affect one and *are affected by one*.

Consider a more sophisticated form of self-awareness that in turn suggests a more sophisticated form of self-understanding. Experiments on 'self-agency' tested whether or not captive, trained rhesus macaques could distinguish a computer icon that they were controlling with a joystick from an icon that was moving randomly on the screen (Couchman 2012, 2015). Couchman et al. define 'self-agency' as the awareness that some actions and consequences are self-generated rather than the result of external forces (Couchman 2012, 2015; Hoffman et al. 2018). These studies found that captive and trained macaques could identify icons that they were controlling with the joystick with the same success rate as humans. Their ability to do so appears to show that they track the difference between events *caused by them* and events that were simply happening, but not caused by them.

Couchman et al. argue that 'self-agency' results from an integration of cognitive information involving sensory-motor cues, prior expectations about the effects of one's actions, and perception of the outcomes of one's actions. The explanation Couchman et al. offer for the evolution of self-agency is that is important for an animal to be able to behave flexibly in uncertain situations in which a habitual response is not adaptive (Couchman et al. 2009). They argue that an awareness of one's own agency affords an animal greater ability to control the outcome of actions in situations that present novel or unfamiliar features. Interestingly, this understanding of why self-agency evolved matches some accounts of the origins of consciousness itself. As DeGrazia reports, Cabanac et al. argue that consciousness arose from the ability to integrate information from multiple senses (internal and external) and to respond flexibly rather than automatically (Cabanac et al. 2009; DeGrazia 2019). While Cabanac et al. argue that such integration and flexibility emerged in amniotes, early land-dwelling mammals, it is at least clear that it is quite developed in the self-agency displayed by rhesus macaques. The possibility of distinguishing between more or less 'automatic' behaviour in animals – more or less ability to control and flexibility in one's response to a stimulus – allows for the recognition of degrees of

animal self-understanding, just as it does for degrees of animal self-awareness, that are indexed to the flexibility and control animals are able to exert over their own behaviour in various domains.

David DeGrazia recently amended his own concept of bodily self-awareness to foreground the additional criterion of agency. *Bodily agential self-awareness*, he argues, is indicated not only by presence of sensations or by automatic pursuit and avoidance behaviour, but by the ability act flexibility in response to stimuli as a result of learning (DeGrazia 2019). This further criterion allows us to avoid the thermostat problem in the case of self-understanding. We should be convinced that the animal already possesses more rudimentary awareness of its own bodily states, for example that it experiences pain, pleasure, hunger, and other forms of bodily sensations. Agential self-awareness should be thought of as developing out of the ability to behave flexibly in response to awareness of such sensations. It too involves the development of expectations about causal patterns, but here these are causal patterns that exist between such subjective senses and the external world. DeGrazia argues that such agential awareness exists across many reptile, mammal, and bird species, and there is some indication that it may also exist in insects (Barron and Klein 2016). He also argues that such agential awareness involves an ability to form some kind of spatio-temporal 'map' of the world (DeGrazia 2019). Wherever we draw the line, such abilities indicate not only the presence of self-awareness, but also of self-understanding, in which an animal has oriented itself within some set of expectations about its own influence on the world.

5 Social Self-Understanding, Social Agency and Social Self-Awareness

The previous section described agential self-understanding as understanding of causal patterns that involve an animal's own actions and the effects of an animal's own actions. A fuller concept of self-understanding can now be given: self-understanding requires more than sentience or taste avoidance behaviour. It requires that an animal can use information about its own actions and capacities for action to perform behaviours that are not merely 'automatic'. Flexible behaviours suggest that animal assess their own capacities and limitations for action in various aways: Can I make it across the gap? Can I catch the prey? Is the prey worth the effort? Self-understanding, then, is present if an animal develops accurate expectations of causal patterns involving its own actions.

With this concept of self-understanding in hand, we can identify additional forms of self-understanding that can arise when animals that live in stable groups and develop forms of complex social interaction with the same individuals over time. In animal groups that persist over time, in which animals recognise one another, there are often social restrictions on what one can and cannot do. These often take the form of aggressive behaviour directed at individuals who pursue a valued resource, or

who occupy subordinate positions, especially those who transgress their 'rank'. There are also aspects of social structure that permit individuals to 'get away with' certain actions that others might not get away with.

There is extensive evidence that non-human animals understand how social relationships work in the groups that they live in. For example, Dorothy Cheney and Robert Seyfarth's pioneering field experiments with baboons (*Papio Ursinus*) used playback experiments to show that baboons track relationships among conspecifics (Cheney and Seyfarth 2008). In other words, baboons understand who is dominate or subordinate to whom, who is related to whom, and which individuals are likely to help one another if a conflict arises. Baboons demonstrated such knowledge in Seyfarth and Cheney's experiments through clear reactions of surprise when hearing various types of vocalisation that break expectations related to the current standing of relationships between individuals in the group (Cheney and Seyfarth 2008). To take one telling example, baboons react more strongly – they show more signs of surprise and even distress – to rank reversals *between* kin groups than *within* kin groups. This makes sense given that rank reversals between kin groups have more potential to disrupt the social organisation and dominance hierarchy in a baboon troop.

In addition to the concept of *bodily agential self-awareness*, David DeGrazia also introduced the concept of *social self-awareness* into philosophical discussion of animal minds. He defines it as 'awareness of oneself as part of a social unit with differing expectations attaching to different positions' (DeGrazia 2009). DeGrazia argues that many group-living animals including baboons, great apes, dolphins, elephants, and wolves and domestic dogs are socially self-aware on the basis that they demonstrate *social understanding* (DeGrazia 2019). In other words, DeGrazia concludes from studies like Cheney and Seyfarth's that baboons, for example, demonstrate social self-awareness because they understand general social dynamics and relationships in their group. This is an important point, but just as with bodily agential self-awareness, it is also crucial here to include the criterion of flexible social *agency*. In other words, social self-awareness does not only involve an understanding of the group structure, but also an ability to use such knowledge to act successfully in the group.

An example from hamadryas baboons (*Papio hamadryas*) that has been cited regularly in philosophical literature is illustrative here (Bermúdez 2007; DeGrazia 2009). Kummer (1982) originally reported in a discussion of tactical deception (an example initially regarded as evidence of 'theory of mind') that an adult female spent 20 min gradually shifting her seating position over a distance of two meters to a place behind a rock where she began to groom a sub-adult male follower of the group (not one ordinarily belonging to the group) – an interaction not tolerated by the adult male (Kummer 1982). The adult male could see her, but not that she was grooming another male. Whiten and Byrne argue that the female *understood* that the harem male leader could not see that she was grooming another male (Whiten and Byrne 1988).

Such behavioural inhibition in the presence of a dominant individual seems to show that the female baboon had expectations of how the social world *works*, and

that the accuracy of these expectations meant either 'getting away' with what she wanted to do, or being the recipient of an aggressive attack. She understood what would happen if the resident adult male were to see her grooming the sub-adult male. These forms of social expectation and self-understanding have also been studied experimentally in captive rhesus macaques (Drea and Wallen 1999). In one experiment, macaques were taught to solve a simple colour-association task to learn the location of boxes baited with peanuts. The monkeys were tested in two social situations: as a complete social group and as a 'split' group, where half of the troop – either the dominant or subordinate matrilines – were removed from the testing area. In both conditions, the dominant individuals retrieved the food from the baited boxes. In contrast, the subordinate individuals retrieved the food correctly only when in the split condition. Because the subordinates performed well in the split condition, their performance in the combined condition suggests that they inhibited expressing their knowledge in the presence of dominant individuals. Again, this behaviour suggests that they had some expectation about what would have happened had they not inhibited their knowledge in this way. Just as understanding that the wind is causing the bag to fly in the air, this form of social understanding is a form of causal understanding of how social interaction works. Yet, like conditioned taste avoidance, it involves expectations about causal patterns involving an animal's own actions.

These examples suggest that we should amend the concept of social *self*-understanding in a way parallel to our amendment of the concept of understanding in general. Baboons might understand how the social world works, but we cannot yet speak of social *self*-understanding until there is evidence that they can flexibly use information about the relationships between others, and their own relationships to others, to perform successful actions. Fortunately, there is also abundant evidence that many animals can do this, but we will continue with examples from baboons to fill in our picture of this highly social and socially intelligent species. In an example drawn from ongoing research at the Tsaobis Baboon Project in Namibia, experiments are performed in which individuals are given the opportunity to explore small novel foods (Carter et al. 2014). Not all individuals are as willing as others to explore let alone consume the novel foods, but some individuals learn quickly the value of the foods. One particular example is demonstrative (Carter, personal correspondence). An adult female, 'Yaoundé', was presented with and quickly ate a slice of apple dyed red. Later, her sub-adult son, 'Okavango', was given the same stimulus. Although he was interested in the apple, picking it up and exploring it, he did not eat it. Yaoundé saw Okavango with the apple and approached him, at which point he turned and walked away from Yaoundé, preventing her from acquiring the food. Yaoundé approached Okavango to groom him, which he allowed. After several minutes of intense grooming, Okavango relaxed to the point that he dropped the apple, at which point Yaoundé snatched up the novel food, ate it and walked away. The observer had the impression that the grooming was a ruse to acquire the apple piece.

Examples like this one show that non-human animals can understand where they 'stand' in relation to other individuals and in relation to the general social dynamics

of the group. They can use this understanding to perform goal-directed, successful behaviours that take advantage of such information. One knows what one can do, and cannot do, by understanding something about the general causal patterns that make up baboon social life. From a metaphysical perspective, the social understanding and social self-understanding of animals are interesting because they do not involve knowledge of a world wholly external to the animal agents themselves. In other words, it is, in part, baboon social understanding and self-understanding that makes the social life of the group unfold in the way it does. There is a feedback loop between expectations of social patterns and the existence of those patterns in way that has been insightfully analysed in philosophical work on social ontology (Haslanger 2013; Hacking 1995).

6 Awareness and Understanding

Inquiry into the nature of animal understanding and self-understanding affords an opportunity to inquire into the relationship between understanding and awareness in general. The relationship between understanding and awareness is a central issue for familiar thought experiments such as John Searle's 'Chinese room' or the Turing test, and other examples involving automata and thermostats. A familiar line of thought is that while thermostats, computers (up to now), or translation rooms might be able to track changes in environmental 'inputs', and deliver the appropriate 'output' in response to changes in 'inputs', they are not aware of their own states, computations, the inputs and outputs themselves, or the causal relationship between inputs and outputs. In Searle's translation box, the idea is that neither the box as a whole, nor any of its parts, understands the meaning of the words it is correctly translating. While this example involves linguistic understanding, the others raise the question of whether systems that respond appropriately to stimuli – in some cases flexibly – can be said to understand the world or themselves if they are not aware of the world or themselves.

Intuitions about the general relationship between understanding and awareness are involved the way we interpret non-human minds. Forms of life can be more like unaware thermostats, but the way they track of features of the world may also approach the awareness and expectation of causal patterns and dependencies required for understanding. It seems possible for a creature to be aware in some way – to have some form of subjective experience or 'what it is like' to be it – and yet not to understand the world. In other words, if primitive subjective 'experience' has the structure of 'white noise' as Peter Godfrey-Smith suggests, it probably does not track causal relationships in a way necessary to develop expectations about such relationships (Godfrey-Smith 2017). Like Cabanac et al., referenced earlier, Godfrey-Smith looks to evolution of sensory-motor feedback loops to find the origins of consciousness. As organisms begin to track not only what they are doing, but how their experience of the world changes as a result of what they do, their experience moves from undifferentiated 'white noise' to more structure and

integration. Godfrey-Smith favours the view that consciousness arose with the cognitive integration of information from both internal and external sources, and that such integration occurred fairly early on in the evolution of animal life.

Such an evolutionary view supports a certain conception of the relationship between awareness and understanding. While we cannot infer the ability to understand from the presence of awareness or some form of subjective experience, the reverse seems to be a justified inference. If an animal indeed understands the world, or itself, then it must be aware of the world, or itself. It seems right, then, to design experiments that test various forms of understanding – such as the experiments on causation with crows, on 'self-agency' in macaques, and on social knowledge in baboons cited above – and to conclude from evidence that animals understand causal patterns in the world and are able to act flexibly on the basis of such understanding that they are aware of the world in various ways. The inference from understanding to awareness, or self-understanding to self-awareness, (but not the reverse inference) is supported by the conception of understanding as a set of expectations involving causal patterns between external events, or between one's own actions and their effects (both internal and external).

Of course, such an inference is not widely accepted by empirical researchers investigating awareness or self-awareness. Baboons, for example, are often thought to lack self-awareness because they have not to date 'passed' the classic mirror test (Carter, personal communication). That is, they do not appear to be able to recognise their reflection in a mirror *as* an image of their own body. Since the 'mirror test' has been taken as the gold-standard test for self-awareness in animals, many have assumed that species lack self-awareness if they do not perform self-directed behaviours in response to their reflection. Animals from diverse taxa appear to 'pass' the test by performing such self-directed behaviours, such as inspecting their bodies or touching marks placed on their bodies (Gallup 1970). Nonetheless, mirror tests have had controversial results, and the findings are often not as clear-cut as they are sometimes presented. For example, chimpanzees are often claimed to 'pass' the mirror test to the exclusion of all other non-human species (Gallup and Anderson 2018); however, while some chimpanzees undeniably 'pass', others do not (Swartz and Evans 1991). This indicates that mirror self-recognition is not a universal trait in chimpanzees, and may even be learned. Further questions arise when we consider that wild chimpanzees do not respond to a mirror in a similar manner to captive individuals (Anderson et al. 2017). At the other end of the scale, small cleaner wrasse have passed the test by rubbing their body against a rock only when it was injected with a dye and placed in front of a mirror, but this has been denied as evidence of self-awareness due to incredulity that fish possess a physiological architecture complex enough to support self-awareness (de Waal 2019; Kohda et al. 2019).

If we bring philosophical reflection on the nature of understanding and scientific work on awareness and self-awareness together, as I suggest we do, then a wider range of experimental evidence might support inferences regarding awareness and self-awareness. For example, experiments on understanding of 'self-agency' in rhesus macaques ought to have the same status as the mirror test in discussions of self-awareness, and studies of social self-understanding involved in behaviours such

as deception or third-party reconciliation in baboons ought to justify inferences of social self-awareness. The inference from understanding to awareness is sound, and this is because understanding requires not only that animals track causal patterns in the world and in the traffic between the world and their own actions, but also that they act successfully and adaptively on the basis of such understanding. Recognition of various forms of understanding across a range of species can help us appreciate a greater variety of evidence that might be available for awareness and self-awareness.

7 Conclusion

This article has defended a notion of self-understanding as an understanding of causal patterns involving the traffic between one's sensations, one's actions, and the effects of one's actions on the outside world. If understanding involves the ability to track causal patterns in the world, and this has been shown in pre-verbal infants and a variety of bird and mammal species, then understanding does not require propositional knowledge, mastery of the concept of causation, or linguistic comprehension. I have argued that self-understanding does not require these capacities either. A conception of self-understanding that does not involve these capacities allows for the possibility that non-human animals can understand themselves. I do not claim to have gone so far as to have proven that self-understanding is present in the examples given here, but merely that they are good candidates for satisfying the conception of self-understanding given here. More empirical work would be required to demonstrate conclusively that animals accurately track causal patterns involving their internal sensations, actions, and the effects of these actions.

In closing, let me gesture briefly to some broader implications of the concepts of understanding and self-understanding given here. One is that animals and pre-verbal children make sense of the world in a 'first-order' manner before they are able to reflect on how they make sense of the world in a 'second-order' manner. Animals may, then, develop expectations about causal patterns without, apparently, the ability to reflect on what or how they understand. The forms of understanding and self-understanding that may exist in non-human animals have likely been shaped through an evolutionary history of causal traffic between subjective responsiveness and what exists in the outside world. Of course, pre-linguistic, 'everyday' forms of understanding and self-understanding are not 'worldviews' in the sense of ideas about the fundamental nature of things, or of what is good and just, nor are they fine-grained, full causal pictures like what we expect in the natural sciences. But, these more lofty and rigorous forms of understanding seem to presuppose the 'everyday' capacities to discern with how things work in the world that this article has aimed to bring into view.

Acknowledgements and Funding I would like to thank Dr. Alecia Carter (UCL) for discussion of these issues and the anonymous reviewers for helpful comments on earlier drafts. I would like to

thank the Templeton World Charity Foundation for funding that allowed for the time to complete this article (TWCF0502).

References

Anderson J, Hubert-Brierre X, McGrew W (2017) Reflections in the rainforest: full-length mirrors facilitate behavioral observations of unhabituated, wild chimpanzees. Primates 58:51–61. https://doi.org/10.1007/s10329-016-0574-7

Barron A, Klein C (2016) What insects can tell us about the origins of consciousness. Proc Natl Acad Sci U S A 113:4900–4908. https://doi.org/10.1073/pnas.1520084113

Baumberger C, Beisbart C, Brun G (2017) What is understanding? An overview of recent debates in epistemology and philosophy of science. In: Grimm S, Baumberger C, Ammon S (eds) Explaining understanding: new perspectives from epistemology and philosophy of science. Routledge, New York, pp 1–34

Bermudez J (1998) The paradox of self-consciousness. MIT Press, Cambridge

Bermúdez J (2007) Thinking without words. Oxford University Press, Oxford

Blaisdell A, Sawa K, Leising K, Waldmann M (2006) Causal reasoning in rats. Science 311:1020–1022. https://doi.org/10.1126/science.1121872

Bures J, Bermudez-Rattoni F, Takashi Y (2002) Conditioned taste aversion: memory of a special kind. Oxford University Press, Oxford

Cabanac M, Cabanac AJ, Parent A (2009) The emergence of consciousness in phylogeny. Behav Brain Res 198:267–272. https://doi.org/10.1016/j.bbr.2008.11.028

Carter AJ, Marshall HH, Heinsohn R, Cowlishaw G (2014) Personality predicts the propensity for social learning in a wild primate. PeerJ 2:e283. https://doi.org/10.7717/peerj.283

Cheney D, Seyfarth R (2008) Baboon metaphysics: the evolution of a social mind. University of Chicago Press, Chicago

Couchman J (2012) Self-agency in rhesus monkeys. Biol Lett 8:39–41. https://doi.org/10.1098/rsbl.2011.0536

Couchman J (2015) Humans and monkeys distinguish between self-generated, opposing, and random actions. Anim Cogn 18:231–238. https://doi.org/10.1007/s10071-014-0792-6

Couchman J, Coutinho M, Beran M, Smith JD (2009) Metacognition is prior. Behav Brain Sci 32:142. https://doi.org/10.1017/S0140525X09000594

DeGrazia D (2009) Self-awareness in animals. In: Lurz R (ed) The philosophy of animal minds. Cambridge University Press, Cambridge, pp 201–217

DeGrazia D (2019) Animal self-awareness: types, distribution, and ethical significance. In: Fischer B (ed) The Routledge handbook of animal ethics. Routledge, New York, pp 71–82

Drea CM, Wallen K (1999) Low-status monkeys "play dumb" when learning in mixed social groups. Proc Natl Acad Sci U S A 96:12965–12969. https://doi.org/10.1073/pnas.96.22.12965

Eriksson N, Wu S, Do C, Kiefer A, Tung J, Mountain J, Hinds D, Francke U (2012) A genetic variant near olfactory receptor genes influences cilantro preference. Flavour 1:22. https://doi.org/10.1186/2044-7248-1-22

Gallup G (1970) Chimpanzees: self-recognition. Science 167:86–87. https://doi.org/10.1126/science.167.3914.86

Gallup G, Anderson JR (2018) The "olfactory mirror" and other recent attempts to demonstrate self-recognition in non-primate species. Behav Process 148:16–19. https://doi.org/10.1016/j.beproc.2017.12.010

Godfrey-Smith P (2017) Other minds: the octopus and the evolution of intelligent life. William Collins, London

Gopnik A, Sobel D, Schulz L, Glymour C (2001) Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. Dev Psychol 37:620–629

Grimm S (2016) Understanding and transparency. In: Baumberger C, Grimm S, Ammon S (eds) Explaining understanding: new perspectives from epistemology and philosophy of science. Routledge, New York, pp 212–229

Hacking I (1995) The looping effects of human kinds. In: Sperber D (ed) Causal cognition: A multidisciplinary debate. Clarendon Press, New York, pp 351–394

Haslanger S (2013) Resisting reality: social construction and social critique. Soc Theor Prac 40(1): 145–152

Hoffman M, Beran M, Washburn D (2018) Rhesus monkeys (Macaca mulatta) remember agency information from past events and integrate this knowledge with spatial and temporal features in working memory. Anim Cogn 21:137–153. https://doi.org/10.1007/s10071-017-1147-x

Jelbert S, Miller R, Schiestl M, Boeckle M, Cheke L, Gray R, Taylor A, Clayton N (2019) New Caledonian crows infer the weight of objects from observing their movements in a breeze. Proc Royal Soc B: Biol Sci 286:20182332. https://doi.org/10.1098/rspb.2018.2332

Kohda M, Hotta T, Takeyama T, Awata S, Tanaka H, Asai J, Jordan A (2019) If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals? PLoS Biol 17:e3000021. https://doi.org/10.1371/journal.pbio.3000021

Kummer H (1982) Social knowledge in free-ranging primates. In: Griffin D (ed) Animal mind - human mind. Springer, Berlin

Kushnir T, Gopnik A (2005) Young children infer causal strength from probabilities and interventions. Psychol Sci 16:678–683. https://doi.org/10.1111/j.1467-9280.2005.01595.x

Parker L, Rana S, Limebeer C (2008) Conditioned nausea in rats: assessment by conditioned disgust reactions, rather than conditioned taste avoidance. Can J Exp Psychol 62:198–209. https://doi.org/10.1037/a0012531

Reilly S, Schachtman T (2008) Conditioned taste aversion: neural and Behavioral processes. Oxford University Press, Oxford

Saxe R, Tenenbaum J, Carey S (2005) Secret agents:inferences about hidden causes by 10- and 12-month-old infants. Psychol Sci 16:995–1001. https://doi.org/10.1111/j.1467-9280.2005.01649.x

Saxe R, Tzelnic T, Carey S (2007) Knowing who dunnit: infants identify the causal agent in an unseen causal interaction. Dev Psychol 43(1):149–158. https://doi.org/10.1037/0012-1649.43.1.149

Swartz K, Evans S (1991) Not all chimpanzees (Pan troglodytes) show self-recognition. Primates 32:483–496. https://doi.org/10.1007/BF02381939

Taylor A, Miller R, Gray R (2012) New Caledonian crows reason about hidden causal agents. Proc Natl Acad Sci U S A 109:16389–16391. https://doi.org/10.1073/pnas.1208724109

du Toit JT, Provenza FD, Nastis A (1991) Conditioned taste aversions: how sick must a ruminant get before it learns about toxicity in foods? Appl Anim Behav Sci 30:35–46. https://doi.org/10.1016/0168-1591(91)90083-A

Völter C, Sentís I, Call J (2016) Great apes and children infer causal relations from patterns of variation and covariation. Cognition 155:30–43. https://doi.org/10.1016/j.cognition.2016.06.009

de Waal F (2019) Fish, mirrors, and a gradualist perspective on self-awareness. PLoS Biol 17: e3000112. https://doi.org/10.1371/journal.pbio.3000112

Whiten A, Byrne RW (1988) Tactical deception in primates. Behav Brain Sci 11:233–244. https://doi.org/10.1017/S0140525X00049682

Zagzebski L (2001) Recovering understanding. In: Steup M (ed) Knowledge, truth, and duty: essays on epistemic justification, responsibility, and virtue. Oxford University Press, Oxford, pp 235–252

A Methodological Response to the Motley Crew Argument: Explaining Cognitive Phenomena Through Enactivism and Ethology

Mark-Oliver Casper and Giuseppe Flavio Artese

Abstract The enactive approach to cognition is presented as an attractive alternative to mainstream paradigms in the cognitive sciences, rejecting notions such as the ones of information processing, representation, and computation. However, notwithstanding the growing interest received in contemporary debates, enactivism is confronted with critical methodological challenges. One of these challenges is the so-called "Motley-Crew Argument." It makes the critical point that if cognition has to be studied as spanning across brains, bodies, and environment, then enactivists automatically rely on a definition of cognition that is too broad and ultimately amenable to rigorous scientific scrutiny. In this text, we pave the way for a methodological answer to this worry and argue for an interdisciplinary connection between biological ethology and enactivism. We show that both approaches share theoretical commitments and that the methodical repertoire of ethology fits the theoretical perspective of enactivism. An ethological case study on risk evaluation in gregarious birds is presented as an example of how a cognitive phenomenon can simultaneously be approached from an enactivist and ethological perspective.

Keywords Enactivism · Ethology · Motley-crew argument · Methodology · Explanation · Action-readiness · Risk evaluation

1 Introduction

The 4E approaches constitute a branch of cognitive science and their main goal is to understand how cognitive phenomena emerge from the interaction between brain, body, and environment. 4E researchers therefore cut across what is usually accepted as "boundaries of cognitive systems" and aim to radically change the ways we think about and study cognitive phenomena. Enactivism is one of the more prominent 4E approaches besides the extended mind approach, the embodiment theory, and the

M.-O. Casper (✉) · G. F. Artese
Institute of Philosophy, University of Kassel, Kassel, Germany
e-mail: mo.casper@uni-kassel.de; Giuseppe.Flavio.Artese@uni-kassel.de

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_3

embedded mind view. However, enactivism is also confronted with methodological challenges. One of those challenges for enactivists is determining suitable types of explanations for their target phenomena. That enactivism is not explicitly committed in this regard is a relevant disadvantage since the lack of methodological transparency leaves empirically working researchers clueless about how to approach specific cognitive phenomena from an 4E perspective. We claim that an important step in handling this issue can be made by drawing a connection between enactivism and ethological research. We contend that, in the face of enactivist premises and basic concepts, explanation procedures of (cognitive) ethologists can be used to meet enactivism's methodological problems – at least in terms of explaining cognitive phenomena. To our knowledge, not much work has been done to study the theoretical commonalities of enactivism and ethology. To close this gap and to support our claim, we detail the challenges and arguments against enactivism at the beginning of Sect. 2. While we present the theoretical groundings of enactivist and ethological research, we argue that there are strong synergies among the premises and concepts used by both approaches. Two aspects enable their linkage: (i) The shared assumption that interaction patterns are dynamically emerging between organisms and the environment, and (ii) the concept of action-readiness that is used to describe why organisms are prone toward a specific response to stimuli. In Sect. 3, the theoretical commonalities of ethology and enactivism are complemented with an ethological case study focusing on cognitive performances in avian organisms, specifically the social constitution of risk evaluation in several species of gregarious birds. These studies make a plausible case for the mutual complementation of both approaches. So far, there have been no systematic proposals to link enactivism with ethology in this way. Hence, this paper is a first step toward connecting both research branches. Further discussions of our provided argumentation will be necessary to explore how deep and tight their connection can be.

2 Enactivism's Methodological Challenges and First Steps to a Full Response

In the last three decades, it has been claimed, with different degrees of radicality, that cognition should be considered an extended, enacted, embodied, and embedded phenomenon. The research branch of cognitive science that revolves around these phenomena has been labeled in various ways, such as "4E approaches" or "situated cognition research." By using these labels, we follow the liberal definition offered by Vogel et al. (2020) as much as many others (Gallagher 2009; Eliasmith 2009), while we do not suggest a total equivalence – neither between the labels nor of the approaches subsumed under this philosophical umbrella term (see, for example, Di Paolo 2009 or Kiverstein and Clark 2009 for a discussion over the differences between enactivism and other 4E approaches). The main goal of situated cognition research is to "flip the script" in the conceptualization of cognitive phenomena. A

substantial fraction of 4E researchers no longer assume, as cognitivists do, that those phenomena are at their core computational, representational, and internally constituted. Instead, the hallmark claim of the 4E approaches is that the entities and processes that constitute cognitive phenomena are decentralized, heterogeneous (there are biological and non-biological ones), and dynamically interacting (Gallagher 2017). Especially the latter point includes the claim that notions such as adaptivity, plasticity, and non-linearity are significant for studies of cognitive systems. Further details of these assumptions, which will be important for the conjunction of enactivism and ethology, will be laid out in Sect. 2.1.

While the idea of situated cognition sparked comprehensive investigations into the nature of cognitive phenomena, the methodological groundings of these investigations seem to be neglected. Although formalizations of dynamical and world-oriented conceptualization of cognitive phenomena have been notably proposed in recent times (Kirchhoff and Kiverstein 2020; Favela 2020; Port and van Gelder 1995), the problem of a "methodological blind spot" of 4E research lingers on. A critical point in this context is the so-called "motley crew argument" (Shapiro 2011). It states that if the hallmark claim of situated cognition research is true, then it is unclear how entities and processes that are supposed to be dynamically distributed over neural, bodily, and environmental factors can be defined and investigated as proper units of research. Additionally, finding scientifically significant regularities in systems that are assumed to be quickly adapting to their surroundings and which exhibit non-linear behavior is a challenging task that must be systematically addressed (Winkel et al. 2009). Tackling this challenge presupposes questions such as "How does cognitive science exactly relate to biological sciences?", "How to identify cognitive systems if their constituents include biological and non-biological entities?", and "How to tell apart constitutive and non-constitutive parts of cognitive systems?". Despite valuable attempts to tackle parts of such concerns (Di Paolo, Buhrmann, Barandiaran 2017), all theories of the situated cognition approach still need to respond to the motley crew argument. Since the individual 4E approaches have their own theoretical commitments, an overall answer to the motley crew argument is unattainable. Each 4E position needs to tackle methodological questions specifically. In this paper, we will particularly focus on enactivism and prepare grounds to handle the motley crew argument.

A consequence of these methodological problems is that 4E researchers are frequently puzzled about how to investigate their objects of investigation meticulously. This puzzlement is broadly noticed and identified by critics as a gap in enactivism's research strategy. More specifically, skeptics claim that enactivists have not established a specific explanatory framework in which cognitive phenomena can be identified, described, and *explained* (Abramova and Slors 2019). A comprehensive procedure on how to explain phenomena under study is still missing despite valuable contributions in this regard (Stepp et al. 2011). We think that such criticism is not sterile and that enactivism indeed needs to confront these methodological issues. To correctly identify cognitive phenomena and explain them by analyzing the constitutive factors behind them, enactivists must supply an explicit methodology showing which explanation types are applicable. "Methodology," in

the context of situated cognition research, refers to the philosophical study of methods already employed within the 4E branch. This includes not just the mere identification of available methods but also their analysis (e.g., in terms of implicit theoretical commitments incurred when applying them) and evaluation (on their efficiency and compatibility). A methodological investigation also requires that enactivists commit themselves to a philosophical position that clarifies the reason behind the applications of their methods. "Methods" are structured proceedings applied in scientific practice for various purposes (developing concepts, issuing theories and models, making observations, gathering data, constructing experimental settings, offering explanations, and providing predictions) (Kaplan 1967/2017, p. 18ff.). While, on some occasions, it has been claimed that enactivism could be considered just a general philosophy of nature (Gallagher 2017), in this context, we follow the more general and shared idea of enactivism as a research program that is supposed to represent an alternative to cognitivism as proposed initially by Varela et al. (1991/2017). If this second option is followed, enactivism, like all good research programs, must be able to offer coherent methodological guidelines. A specific point might help to elaborate this point further.

The 4E debate as a whole includes several explanation types such as mechanistic (Craver and Darden 2013), dynamical (Lamb and Chemero 2014), normative (Satne 2015; Casper 2019), teleological/etiological (Garson 2011; Millikan 1984), ethological (Jamieson and Bekoff 1992), phenomenologically-inspired (Rietveld and Kiverstein 2014; Gallagher 2005), and functional (Weiskopf 2011; Wheeler 2010) explanations. In the following sections, we focus on autopoietic enactivism and confront its lack of commitment on the (in-)compatibility of different explanation types and employed explanation procedures. We argue that enactivists need to choose specific explanation types and accept that such a choice excludes using other types of explanation. Furthermore, we claim that ethological explanations should be incorporated into a comprehensive enactivist research agenda – including the comparative study of cognitive phenomena across different species. The aim is to develop a proper explanatory framework that works empirically with the mind-life continuity thesis, which implies that cognitive competencies are an inherent aspect of living beings (Thompson 2007). By doing so, enactivism can deal with the motley crew argument since ethological approaches can handle whole catalogs of entities and processes while identifying the relevant ones for investigating behavioral profiles over populations (or their individuals).

2.1 Enactivism: Basic Ideas and Concepts

Among the 4E approaches, the enactive approach to cognition can be considered the most radical. In contrast to the extended mind theory or weaker embodiment theses, enactivism has not been presented as a simple revision of orthodox cognitive science but as a proper alternative. Enactivism rejects the brain-computer metaphor and the overall notion of computation (see Casper and Artese 2020) and especially argues

against the claim that cognitive systems require an input-output conversion of information that is couched in representational terms. Importantly, enactivism has been subject to a cascade of theoretical differentiations. At the moment, it is possible to distinguish at least three different varieties of enactivism. Following Ward et al. (2017), the different "families" of enactivism can be identified as (i) Autopoietic Enactivism, (ii) Sensorimotor Enactivism, and (iii) Radical Enactivism. Whether these ramifications are compatible with each other is an open subject of discussion that goes beyond the scope of the paper (but see Thompson 2018, Hutto 2005, Noë 2021). However, since the sensorimotor approach is mostly focused on explaining sensorimotor regularities in human perception and radical enactivism is more dedicated to providing theoretical parsimony through significant contributions in terminological debates, we will focus on what is generally defined as autopoietic enactivism. This choice is motivated since both – autopoietic enactivism, as developed by Di Paolo (2005, 2018), and ethology – are particularly focused on phenomena emerging from the interaction between an organism with its environment. Secondly, autopoietic enactivism represents, among the different varieties of enactivism, a version that explicitly aims at establishing a scientific and coherent research program in the cognitive sciences. In contrast, the sensorimotor approach has been primarily focused on very restricted domains of applications.[1] At the same time, radical enactivism seems more directed at providing a broad theoretical framework in which a new science of the mind is purged from its reductionistic and representationalist commitments.

Autopoietic enactivism (that from now on will be simply addressed as enactivism) is, at the same time, a theory of life and cognition. It prominently defends a strong continuity between both that is expressed with the mind-life continuity thesis (Thompson 2007). The enactive approach can be characterized by five principles (ibid., 2007). The first principle is that "living beings are autonomous agents that actively generate and maintain themselves, and thereby also enact or bring forth their own cognitive domains" (ibid., p. 13). This idea implies the concept of autopoiesis (Varela et al. 1974). Enactivism assumes that, as autopoietic beings, organisms should be thought of as constituted by a circular set of processes that guarantee the organism's ability to maintain its identity. Because of their self-referentiality, organisms can be classified as autonomous – they establish an individual perspective on the environment. To not disintegrate and maintain their complex set of circular operations, the organism finds itself in a continuous state of precariousness. The maintenance of the organism's autopoietic processes is possible through the exchange of the right kind of gradients with the environment and the avoidance of factors that could be harmful – otherwise, its integrity would be compromised. A significant feature of autopoietic organisms is their capacity to be

[1] Here, with "sensorimotor approach," we are just referring to the work of O'Regan and Noë (2001) which can be mostly related to a theory of sensory consciousness. We are aware that the tools of sensorimotor enactivism have been employed and implemented into autopoietic enactivism and extended to a large number of cognitive domains (see Di Paolo et al. 2017).

adaptive. Adaptivity needs to be understood in this context as the capacity of a system to actively regulate its relation with the environment and itself "with respect to the boundaries of its own viability" (Di Paolo 2005, p. 8). The latter has been accepted as essential by most enactivists since it is theoretically necessary to think of organisms to have a degree of preferences towards environmental stimuli.

A classic example of adaptive behavior in autopoietic organisms is the one of an E. Coli bacterium swimming towards a higher concentration of sugar and ignoring regions with lower concentrations. In this sense, by being "autopoietic" and "adaptive," a system classifies as autonomous if they fulfill these two criteria. In light of the scope and aim of the paper, similar considerations can be applied to more or less sophisticated animals. An interesting case is represented by earthworms, whose behavior has been intensively documented by Darwin (1881). Even lacking a nervous system and complex sensory organs, earthworms can show a high degree of selectivity towards the environment, particularly in how they dig their burrows. Earthworms take care of the composition and temperature of the soil, dig the shape of their burrows in such a way to support and boost their capacity to produce rapid movements when required, and are highly selective in the kinds of materials used for lining or for building their basketlike nests. Earthworms make sense of the environment by following their biological norms, expressed by being selective towards the gradients of their surroundings. If the enactivist strategy is followed, then *biological norms* allow us to speak about an individual cognitive domain or perspective established on the environment, in this case, by bacteria or earthworms.

The second criterion is that *the contribution of neural activity to cognitive processes can only be grasped if the brain is considered an autonomous dynamic system*. From the enactive perspective, the role assigned to the brain is not the one of calculating possible solutions to some environmental problem previously reconstructed from the raw stimulations of the organism's receptors. What the brain does, enactivists argue, is to generate and maintain meaning. Far from being considered a detached unit of analysis, the brain is described in dynamic terms and as inseparable from a lived body embedded within an environment (Cosmelli and Thompson 2010; Thompson and Cosmelli 2011). The central nervous system is primarily considered an organ whose activity is highly interconnected, dynamic, and constrained by the history of interactions of a bodily subject (Dotov 2014; Van Orden et al. 2012). These considerations imply that the central nervous system is not a representational device that stores information but an organ for action that should be understood non-representationally.

The third criterion proposed is that *cognition consists in the exercise of (skillful) know-how in embodied action.* As has been mentioned, the role of action is crucial for the enactive approach and, consequently, the full history of interactions of individual perceivers. As Thompson (2007) and, more recently, Di Paolo, Buhrmann, and Barandiaran (2017) forcefully argued, cognitive architectures and capacities emerge from the continuous loops of action and perception at the basis of the everyday life of a given animal. As prominent work in dynamical neuroscience demonstrates, the meaning of brain activity depends in a non-trivial way on the actions that a bodily creature has performed for short and extended periods in its

environment (Freeman 2000). At the same time, even if, as Varela noticed, "the state of activity of sensors is brought about most typically by the organism's motion" (1997, p. 82), individual and momentary action and perception loops can only constrain but not fully determine brain activity. Varela argued that the endogenous activity of the brain could be appreciated by focusing on the role played by interneurons. These networks of neurons can be seen as aimed at constraining the activity of both sensors and effectors on different temporal scales and, at the same time, the overall processes necessary for maintaining the organism's self-referential identity. Di Paolo, Buhrmann, and Barandiaran developed a detailed account of what they defined as "sensorimotor schemes" to make sense of the emphasis offloaded by Varela on the organisms' endogenous activity. This notion is intended as a mesoscopic level of description of the "reusable, interlocking, organizing, sets of coordination patterns between body and environment" (2017, p. 81) that agents can enact in virtue of their histories of interactions. From the enactive perspective, these schemes can be seen as multiple and complex action-readiness states. The combination of different sensorimotor schemes allows an individual to deal with environmental situations in such a way that would not be possible if his activity were reduced only to feedback loops between sensors and effectors. Different motor schemes might reinforce or inhibit some others in such a way as to create proper networks of significance. These groups of schemes have been defined by Varela (1997, 1999), later by Di Paolo et al. (2017) and Kiverstein and Rietveld (2018) through the concepts of microidentities and microworlds. Following Di Paolo and colleagues, microworlds are realms of significance that define the "frame" of a situation in which sensorimotor schemes are enacted (2017, p. 167). Parallel to the enactment of its microworlds, the organism also develops its states of action-readiness that allow it to maintain a grip and act on the situation.

The fourth criterion has a strong ontological flavor. We mention it for the sake of completeness. However, it will not be relevant in the further sections. This criterion states that the concept of enaction presupposes the *existence of a subject-dependent environment (instead of a pre-given reality) that is mutually co-determinate by physical aspects of realities as much as from the autonomous identity of the perceiver*. This point, which represents the ontological foundations of enactivism and is supposed to differentiate it from its cognitivist rivalries, is the phenomenological consequences of the concept of *autonomy* and the relationship of mutual determination between organism and environment. How the environment appears to a given creature depends on the autonomous processes as much as on the physical constitution of the world. Thus, the consequence of autonomy is that the world is not perceived as an objective subject-independent reality but rather strongly differs based on the individual organism and its history of interactions.

The last and fifth criterion states that enactivism is an experience-centered approach. In contrast to mainstream cognitive science, in which subjectivity is often considered epiphenomenal or an obstacle to overcome, the enactive approach focuses on such subjective qualities. Subjectivity, and notions usually associated with an eventual mental domain, are generally pictured as a guiding force in the explanation of mental phenomena. It will never be emphasized enough that

enactivism strongly resists attempts to reduce the phenomenal domain into the physical but argues for its emergence from the autonomous processes of an organism situated in the environment.

2.2 The Ethological Approach to Behavior

We argue that enactivism, as specified so far, can be at least partly connected with ethological premises and research strategies. Such a connection enables enactivists to confront the above-mentioned methodological concerns held against them. On the one hand, the link between the two approaches seems unproblematic since enactivists assume that cognitive phenomena are constituted by the competence of organisms to organize their reactions to environmental stimuli autonomously. This competence is assumed to be an essential feature of biological systems. Cognitive phenomena are therefore conceived of as biological phenomena. Since ethology is defined as the study of the biology of behavior (Immelmann 1983), its research can be used to analyze what, from an enactivist perspective, is the constitutive basis of cognition: interaction patterns exhibited by autonomous organisms. Thus, we assume that ethological analysis can represent a scientific formalization of the activities of different species necessary to enact and which structurally couple them to their meaningful environments. Therefore, we assume that such analyses are directly in line with the enactive assumption of cognition as the enactment or capacity to bring forth meaningful surroundings in virtue of the agent's sensorimotor capacities.

On the other hand, cognitive phenomena are not the primary focus of ethological research. The study of behavioral profiles in animals is not necessarily aimed at their cognitive states. If enactivists employ an ethological approach, it apparently needs to be a form of *cognitive* ethology. This section presents the basic assumptions of ethology and cognitive ethology, emphasizes the continuity between both, and highlights the possible theoretical links between ethological approaches and enactivism. In the following, we define ethology in a possible but also beneficial way for the paper's overall goal to align enactivism with ethology. Divergent understandings are possible.

Ethology is a subdiscipline of biology that has the behaviors of animals as individuals or populations as its subject. Ethological investigations focus on whole patterns of behavioral performances under the most natural conditions (Lehner 1979). Being close to such natural conditions is relevant for understanding *why* behavioral profiles might have developed and stabilized (Bekoff 1999, p. 371). A change of conditions for experimental purposes might distort the phenomenon under study. Studying behavioral patterns means investigating the (organismic as well as environmental) requirements under which unconditioned (innate) and conditioned (learned) forms of behavior are triggered. It further includes analyzing the threshold and threshold change of behavioral responses to a specific stimulus, while the responses can be complex patterns that go beyond mere reflexes. The term

"threshold change" refers to the phenomenon that a specific *action-readiness* and its changes depend on various aspects such as consumed energy resources, high repetition, or modification of ecological influences (Immelmann 1983). To identify, describe, and explain (the constituents of) action-readiness is the most important focus for ethology.

While a description of an animal's behavior is preferred, which is as theoretically neutral as possible, the goal of ethology is to predict and explain the behavior of individuals, groups, or whole species in specific situations. While the description of behavior is strictly separated from the explanation of behavior (Burghardt 1973), ethological studies have roughly four major explanatory foci, which are (i) the function of a behavior (function is understood as the "survival *value*" of behavior), (ii) the ontogeny (the individual development) of (inter-)action patterns, (iii) the phylogeny (or the evolution) of behavioral profiles, and (iv) the mechanisms of behavior which are internal processes of an organism that are relevant for a given behavior to be manifested (Tinbergen 1963; Bateson and Laland 2013). In sum, ethology investigates the endogenous and exogenous triggers of behavior. General behavioral regularities shall be demonstrated and understood.

Therefore, it is crucial to draw a segmentation. *Descriptive Ethology* focuses on a species' morphological aspects (Immelmann 1983). This includes the description of physiological structures of the body, such as the form, structure, relation, and mutual stabilization of bodily tissues. Descriptive ethologists also name behavioral profiles, break these patterns into recognizable units, and compile an "ethogram." An ethogram can be defined as a catalog of basic performances that, if combined, can also constitute complex performances (Lehner 1979). *Experimental Ethology* designs experimental settings in which the elicitation of behavior can be controlled. Triggers of behavior can be identified, described, and predicted. This procedure includes laboratory-based research and field studies, which are the preferred experimental settings. *Comparative Ethology* explicates and compares the behavioral patterns across species to understand the various solutions of animals to common problems and the major principles of behavioral determination.

Finally, Cognitive Ethology extends the classical subject of ethology to animals' cognitive and conscious states. It includes phenomena such as thought processes, rational beliefs, and *consciousness* (Bekoff 1999, p. 371). Since all these phenomena are not supposed to be reduced to organism-intern mechanisms but considered as results of an interplay of various levels (including survival values of behavior, subjective qualities of experience, and environmental structures), it links well with the fifth criteria of enactivism – to consider not just cognitive states and processes but possible experiences of animals as well. Importantly, since we just mentioned mechanisms as possible entities for ethological research, we need to add that a connection of ethology and enactivism will integrate mechanistic positions in the 4E context. Whether mechanistic approaches can be employed in 4E research is a hot and open debate. While most enactivists do not particularly rely upon or reject mechanistic explanations (see Thompson 2007; Di Paolo et al. 2017), others seem open to a possible enactive reinterpretation of the notion of mechanism (see Abramova and Slors 2019). We think that both interpretations can accommodate

ethological explanations within their research agenda. For example, when applied in the concrete contexts of ethological studies, the notion of mechanism refers to the biological substrate of a given animal relevant to triggering a determinate behavior. Enactivist researchers can thus, based on their theoretical assumptions, interpret the reference to such substrates as actual mechanisms as theorized by Gallagher (2018) or as biological factors necessary to make sense of the organism-environment coupling without relying on the heuristic strategies of mechanistic approaches (see Silberstein and Chemero 2013).

The aim of cognitive ethology is thus to license inferences from observable performances of populations (or their individuals) to their cognitive capacities. This is especially the case in highly functional behavior (behavior that ensures survival) such as food-hoarding and – hiding (Morand-Ferron et al. 2016). Populations of the same species confronted with varying selection pressures must evolve special and different behavioral profiles. For example, populations that need to withstand, e.g., more prolonged and colder seasons, might have better spatial memory as advanced food-hoarding and -retrieval is a decisive factor for survival in such populations. Generally, cognitive ethology does not diverge radically from other ethological approaches and their methods but provides a modified interpretation procedure of results.

Two points are essential in this context: (i) The characterization of behavioral performances as a manifestation of cognitive competencies is supported where cognitive traits are most likely to explain the positive effect on survival – as in the case of spatial memory that can be necessary for specific forms of advanced food-hoarding. This point strongly refers to the "inference to the best explanation" strategy explicitly endorsed by cognitive ethologists (Allen 1998). (ii) Selection pressures can be responsible for asymmetrically distributed power of cognitive capacities over different populations of the same species. This issue of the phylogenetic distribution of cognitive phenomena within and between different species is part of cognitive ethology (this turns cognitive ethology into comparative ethology). In addition, the research focus on cognitive ethology overlaps with studies on the ecology of cognition (Mettke-Hofmann 2014; Hutchins 2010). Both types of studies investigate, e.g., the evolutionary induced effects on cognitive developments such as the population's geographic position, time spent under local conditions, or interactions with conspecifics. The extension of inferential licenses from descriptions of behavioral profiles to cognitive capacities is therefore not an exotic peculiarity motivated by speculative ethologists trying to push the boundaries of their discipline. Recently, in the 4E context, different disciplines cluster to comprehend the constitutive, ecological forces on cognition's development while cognitive ethology is directed towards exactly those forces.

In addition to these ethological premises, the question needs to be asked how explanations in ethology are supposed to work. They address and intertwine why- and how-questions simultaneously. Why-questions aim at functional characterizations of behavior. "Functional characterization" in this context means understanding the contribution of a specific interaction pattern to an organism's survival. It is, therefore, decidedly different from "functionalist approaches" in the philosophy of

mind, whose aim is to understand cognitive states as constituted by the functional role they play in interaction with other cognitive states.[2] With how-questions, on the other hand, researchers explore the organism-intern variables of a given behavior. Ethology, at the same time, considers the evolutionary pressures that change the workings and relevance of those variables. Hence, ethological explanations are historical since the time under certain living conditions (including individual life histories or learning histories) is an important factor when explaining behavioral profiles (Blumstein 2006). It also makes such explanations normative because they factor in a population's (good or bad) adaptation to its (local and perhaps temporary) living conditions. Functional explanations involve the evaluation of a population's overall fitness (Morand-Ferron et al. 2016). The consideration of norms makes ethological explanations a perfect match with the first criteria of enactivism that emphasizes biological normativity as a significant aspect of cognition research. Even more in line with 4E-friendly approaches, ethological studies resemble the main idea of Jamesian functionalist psychology, which is generally considered a predecessor of radical 4E approaches in the cognitive sciences (Chemero 2013). Again, the notion of function needs to be confused neither with the functionalist approach in philosophy of mind nor with the cognitivist notion of function. It is understood as a developed adaption to the environment. This makes ethological explanations unique as they (i) analyze cognition's situatedness (in the wild or in the lab), (ii) consider the variation of life histories between populations (or individuals), (iii) use normative vocabulary to include overall fitness evaluation, and (iv) consider organismic factors. By intertwining why- and how-questions, (cognitive) ethology reaches for multi-level explanations that imply the micro-level of organism-intern analysis of internal variables, the meso-level of an organism's behavior plus adaptive functions, and the macro-level of social interactions and environmental forces. That ethological investigations consider the dynamical relations between different levels make them fit with the second criteria of enactivism that embeds the internal processes of organisms, such as brain and bodily processes, in a broader context.

2.3 Conceptual Common Ground: The Notion of Action-Readiness

Contemporary enactivism fiercely argues that cognitive processes are realized within and through the environment (Barandiaran 2017). This emphasis supports our claim that the methods and results of ethological studies can support and be seen as a resource for enactive approaches to cognition.

[2] In the context of cognitive ethology, it is useful to understand the "term" functional in a similar fashion of the tradition of William James in which cognitive capacities, habits, emotions and behavior more generally are studied in terms of their evolutionary or temporary adaptation towards an environment in flux (see Käufer and Chemero 2021).

In this section, we evaluate the usage of specific concepts employed by ethological studies and enactivism. Particularly, we focus on a notion central to the explanatory work of ethologists that simultaneously became important in many recent enactivist writings. It is the concept of *action-readiness,* a notion that is of essential importance for ethologists to describe the tendencies of an animal to act with respect to a particular environmental event (Immelmann 1983). The latter is often also used to indicate the animal drive or its disposition to act in virtue of an intertwinement of organism-related biological processes and environmental triggers. What is particularly interesting about action-readiness is that it presupposes a coupling of perception and action for executing a given form of behavior. If this interpretation of action-readiness is correct, then the ethological emphasis on the mutual determination of organism-related and environmental factors supports the direction of action-oriented approaches in cognitive science. Moreover, by presupposing a diachronic and relational understanding of how actions are manifested in the animal-environment system, the concept of action-readiness can be used to cover and explain several cognitive phenomena as generally assumed by different 4E theories and in particular from non-representationalist and embodied-friendly approaches (see Frijda 2007; Varela 1999; Kiverstein and Miller 2015).

Enactivists already apply these definitions to relatively simple organisms to explain how the continuous cycles of perception and action shape and allow the organism's neural architecture to develop specific states of action-readiness (Varela 1999). Take the case of *Aplysia,* a water mollusk (often known for being one of the largest sea slugs) with a very simple nervous system constituted more or less of about a few thousand neurons. When the Aplysia's siphon touches or is touched by a surface, it contracts its gill. These contractions, which are defined as gill-withdrawal reactions, are essential in the life of the Aplysia and are unlikely to be understood as mere blind and hard-wired reactions. Instead, such contractions should be seen as mediated by a reflex-arc and thus as a unitary psychophysical process in which "sensory stimulus, central connections, and motor responses shall be viewed, not as separate and complete entities in themselves, but as divisions of labor, functioning factors, within the single concrete whole now designated the reflex arc" (Dewey 1896, p. 358). The concept of reflex arc implies that action-perception loops are modulated by the organism's history as much as from the concrete context in which its same action-perception loops are embedded. In the case of Aplysia, empirical studies demonstrate that a large portion of the nervous system of this animal is active during these contractions. Following Varela and the studies he relied on (see Carew et al. 1983 for a review of several studies demonstrating that the Aplysia californica can exhibit several forms of habituations and sensitizations toward meaningful stimuli), the reason the reflex-arc of the Aplysia works the way it does is the presence of basic forms of memory and learning in these creatures. These networks of neurons become active in a highly coordinated and mutually influential manner for a matter of a few seconds. Varela's interpretation is that "the neurons of even this invertebrate ganglion must be conceived as a network of overlapping ensembles which arise in various coherent configurations depending on the animal's context" (Varela 1999, p. 42).

A plausible enactivist argumentation is that the states of action-readiness are equally determined by the organism's action-perception loops, the historicity of the

organism (both developmentally and ontogenetically), and by its coupling with a given environmental context (Varela 1991). We claim that, if put in these terms, *action-readiness* tightly connects enactivist and ethological studies. As mentioned in the previous section, ethologists take very seriously the idea that behavior is not a mere reflex but an actual complex phenomenon involving organismic and environmental constraints and that the latter cannot be properly understood if not as a product of the individual and evolutionary history of an organism inherently coupled to the environment. Immelmann, for example, interchangeably uses the notions of "tendency or readiness to act," the one of "motivation," or the one of "drive." In ethology, these two concepts represent the tendencies of an organism, in virtue of environmental and biological organismic factors, to behave in one way or another (Toates 1986). Similarly, Immelmann, in his "introduction to ethology" (1983), argues that states of action readiness are the resultant of many of several internal (organismic) and external (environmental factors). The main factors in this regard are:

1. *Internal sensory stimuli*. From an ethological point, and thus from the perspective of explanation types that are focused on the behavioral dimension, it is very likely that the action-readiness states of an animal observed in its natural environment are often triggered by the states of hunger or thirst experienced.
2. *Motivating key stimuli*. This category refers to external stimuli in a broad sense. It can include very different factors such as olfactory stimuli, the presence of certain colors in the environment, parents' warning call, higher concentrations of chemicals or oxygen in a given portion of the environment, light, and visual cues. The list could be longer, but the overall idea is that environmental stimuli non-trivially trigger, determine, modulate, or orientate action-readiness states. Notably, the same kind of reasoning needs to be applied to the presence of certain stimuli and their absence.
3. *Hormones.* The presence of hormones produced across the organism's body is necessary to consider if an ethological explanation of drives and actions is wanted. It seems quite spontaneous to think that hormonal changes are at the core of instinctive behaviors in the animal kingdom (see, for example, the role of cortisol in flight-or-fight behavior). Interestingly, several enactivists and embodied theorists have emphasized the necessity of considering hormones in shaping behavior and cognitive processes (see Bower and Gallagher 2013; Colombetti and Zavala 2019). Similarly, it has been recently proposed that ethology can be a source of inspiration for action-oriented, autonomous, and Rodney Brook-inspired robotics. More specifically, Avila-Garcia and Cañamero (2005) argue for the implementation of components that simulate the role of hormonal modulations in perception and action-preparation generally studied by ethologists.
4. *Endogenous Rhythms.* This category refers to the so-called bio-clocks and regular patterns observable in the behavior of an incredibly high number of species whose actions are triggered by cues specifically associated with nocturnal, daily, or crepuscular periods. Interestingly, the relation between different types of biological rhythms and embodied cognition has been recently pointed out by Fuchs (2018). Particularly, Fuchs suggested that the role of cycle rhythms has

been largely downplayed in contemporary debates that aim to characterize how cognitive and perceptual phenomena manifest in subjective experience. Some of the examples Fuchs indicated can be seen in the periodicity of heart rhythms, respirations, sleep-wake cycles, and hormone secretion, among others).

5. *State of Maturation*. The states of action-readiness can differ across organisms of the same species that can be observed in the different life stages.
6. Previous History of a Behavior. Very much in line with the previous discussion initiated by Varela, Immelmann argues that it is essential for ethologists to consider the behavior of an animal from a historical point of view in such a way to take into consideration how often and to which degree past interactions have reinforced a specific action.
7. *Autonomous Production of an Excitatory Potential in the Central Nervous System*. This idea shows another important similarity with the discussion initiated by Varela (1991) and developed later on by Thompson (2007), in which the organisms' activity is not purely determined by the here and now of the situation but also by the endogenous and self-produced activity of the nervous system.

In ethological research, emphasis is put on the states of action-readiness which are understood in a very synergistic manner comparable with ideas promoted in the enactive research program. As argued above, there are no good reasons to exclude the possibilities of theoretical affinities between enactivism and certain branches of ethology and that the latter can be turned into a resource for enactivists.

3 Ethology's Options to Render "Motley Crews" Scientifically Accessible

As mentioned, enactivism is confronted with methodological criticisms such as the motley crew argument or the lack of commitment to explanatory frameworks. While the former is about the difficulty of 4E theories to identify, describe and track all the relevant entities and processes that make up a cognitive system, the latter critical issue points out that no explanation procedure is defined to which enactivists explicitly subscribe. Ethological explanations, as presented above, might be a solution to these methodological concerns. This type of explanation considers organism-intern processes which lead to behavior (micro-level), similarly extends its focus to the whole animal (meso-level), and includes ecological conditions under which its interaction patterns develop (macro-level). Moreover, several entities are simultaneously tracked to understand behavioral and, in the case of cognitive ethology, cognitive phenomena. This focus on the multi-level distribution of relevant behavioral features and constraints matches the enactivist perspective on cognitive systems as distributed over decentralized, heterogeneous, and dynamically interacting entities. Hence, ethological explanations offer strategies to respond to the motley crew argument – if they are connected to enactivism.

Below, we detail the match between enactivism and ethology by presenting a case study and a meta-analysis of ethological research. More specifically, ethologists such as Morelli et al. (2019) study fleeing behavior in birds to understand under which conditions their situational risk evaluation changes. These studies include considerations of local selection pressures such as confinement of space or social group affiliation. We argue that the *modus operandi* of such ethological studies can function as a possible methodological grounding of enactivist research. Finally, to increase the credibility of the Morelli et al. study and, therefore, the overall connection between enactivism and ethology, we present a comprehensive meta-analysis of similar empirical investigations focusing on animal fleeing behavior.

3.1 A Case Study of Risk Evaluation in Gregarious Birds

Several animal species are gregarious. This means that their individuals strongly tend to cluster in groups. In the case of birds, these groups are known as flocks. Flocks have several functions for the individuals joining them. Living in flocks guarantees a lower probability of being caught by predators (Lima 1990), and individuals can invest more time in other activities such as finding resources (Lima and Dill 1990). Being part of a flock influences the selection pressure exerted over organisms. It modifies the predator-prey situation and affects the *risk evaluation* of individuals. Flock size is an important variable that shapes the organism's fleeing behavior and perception of possible threats.

The Flight Initiation Distance (FID) is a significant aspect of fleeing behavior. It is defined as the distance at which an individual takes flight (flees) due to an approaching predator or perceived risk. Morelli et al. (2019) hypothesized that if flock size influences fleeing behavior, then flock size should impact FID. Their study investigated 23 species of gregarious birds located in eight different European countries with 5783 observations. For each species, they provide at least ten observations. Let us note that by including different species of gregarious birds, Morelli et al. also aimed to investigate the phylogenetic distribution of a specific form of risk evaluation in extant gregarious birds; more specifically, they look for behavioral invariances across species. As mentioned above, the phylogenetic distribution of behavioral profiles is a relevant issue in ethological, ecological, and comparative studies of a population's performances. The experiment considered several factors that, as shown by previous research, impact FID, such as the latitude, types of habitat (urban or rural area), diet or foraging strategies, age of individuals, and seasonality (phenological windows). Similarly, body mass is considered as larger birds require more time to get airborne, and therefore they are most likely to exhibit increased FID.

FID was measured "as the distance between the [approaching] observer and the point where the individual bird began to flee" (Morelli et al. 2019, p. 6099). Individual flocks were approached when their members were not vigilant (e.g., roosting or preening), not on a breeding site (in this case, increased vigilance is

assumed that might affect FID), and when they were not occupying an artificial feeding site (e.g., rubbish dumps, those are excluded by the experimenters to avoid possible effects by the abundance of available food). FID was measured only in birds that were standing on the ground. Only single-species flocks were the object of the study. The experimenters approached a flock in a straight line with a constant speed of 0.5 m/s. FID was specified by external observers that relied on binoculars. The studies show that FID positively correlates with body mass, generally decreases in urban habitats and that FID for gregarious birds shows a strong phylogenetic signal. The latter means that the relative FID of individuals in a flock (in rural and urban areas and across species) is positively related to flock size. The main result is that FID increases with flock size in European gregarious bird species.

Morelli et al. explicitly use the intentional characterization of fleeing behavior and its triggers *before* presenting the experimental setup, the results, and their discussions. According to them, birds are "making decisions" whether to escape or stay in a situation of "perceived threats." They infer that members of flocks reach an "escape decision" in certain situations while other species might "take more risk." In this context, the claim that different bird species perceive threats differently, and make different escape decisions, is justified in virtue of the variance of FID in the populations' behavioral profile. Analyzing under which conditions FID changes in populations or individuals is to investigate their *risk evaluation as a cognitive capacity* that is their fleeing performances. The lower the FID, the less risky a situation is evaluated. The higher the FID, the riskier a situation is evaluated (such inferences are only possible if other variables are controlled, such as habitat, body mass, species, or flock size). These intentional characterizations are avoided by Morelli and colleagues while presenting and discussing their results. They merely assume that their "results support the role of sociality in risk-taking behavior" (Morelli et al. 2019, p. 6102).

At the same time, their results can be interpreted in such a way to support an enactivist interpretation of the issue. The study of Morelli et al. focuses, we claim, on a cognitive capacity: *risk evaluation* in specific gregarious birds. As presented above, enactivists assume that cognitive capacities are constituted by interaction patterns between organism and environment (including conspecifics). More specifically, let us stress the third criteria of enactivism: *know-how in embodied action*. The third principle allows us to understand the change in FID due to flock size as a change of what constitutes the bird's capacity to evaluate risks. The fleeing performance of individuals changes according to the number of conspecifics in their flock. Risk evaluation thus appears to be behaviorally enslaved. Individual birds are forced to behave differently once they are part of a flock. Hence, the cognitive capacity of risk evaluation in *gregarious* birds studied by Morelli et al. is constituted by the flock's overall performance and is therefore socially enacted. To put the same state of affairs differently and in more ethological terms: Fleeing behavior in gregarious birds changes since the overall flock and the individual birds mutually modulate each other's action-readiness.

This social organization of risk evaluation can be explained *without* the necessity of individuals picking up information about the situation and its risk values. Internal

information processing of threats is not happening in any flock members (there might be differences between individuals in this regard if we consider their localization – individuals at the center of the flock might differ in risk-related action-readiness compared to those at the periphery of the flock). We argue that enactivists should see this change of FID in correlation to flock size as a case of the social constitution of risk evaluation in birds – and hence as a cognitive capacity that is enacted by social interaction. If enactivists combine their theoretical commitments with ethological research, then the motley crew argument is at least partly handled. As Morelli et al. control variables such as latitude, habitat, foraging strategy, age, seasonality, body mass, flock size and other aspects, they determine the ecological setting of cognition and the entities and processes relevant for the phenomenon to be investigated. These are precisely the entities and processes enactivists need to identify and track if they also want to work empirically on cognitive phenomena. Thereby, methodological concerns about enactivism can decrease to the point at which ethologists are optimistic in identifying relevant ecological conditions for cognitive and behavioral profiles. The second case study will corroborate the connection between enactivism and ethology.

3.2 Further Ethological Support: A Distributed Network of Constituents for Flight Initiation

We argue that the case study by Morelli et al. (2019) is an excellent first instance of how ethological research can be linked to enactivist cognition theory. However, more should be provided to extend the basis for evaluating a possible connection between enactivism and ethology. To strengthen our position despite the low number of case studies we can present due to limited space, we concentrate in this section on a meta-analysis and review of risk assessment in animals (Stankowich and Blumstein 2005) while several other studies could be subject (Fernández-Juricic et al. 2002; Yasué 2005; Burger et al. 2010; Glover et al. 2011; Samia et al. 2017). Also, Stankowitch and Blumstein propose intentional characterizations of animals' fleeing behavior and rely on notions such as "rapid assessment and decision-making" (Stankowich and Blumstein 2005, p. 2630), "perception of risk," "complex decision making," and "assessing risk" (ibid., p. 2632). Like Morelli et al. (2019), such formulations are less used in the context of result presentation and the discussion of findings. However, we take these formulations as hints that, from a 4E perspective, ethological considerations can, with the right amount of theoretical work, be turned into cognitive ethology and thereby connected with enactivism.

Stankowitch and Blumstein present FID as "an excellent metric" (Stankowich and Blumstein 2005, p. 2627), with which a significant aspect of animal fleeing behavior can be quantified. They offer a group of four categories relevant to (the change of) FID in specific populations: "*aspects of the predator, physical condition of the organism, environmental factors, and effects of experiencing and learning*" (Stankowich and Blumstein 2005, p. 2630). In these groups, further aspects beyond

the list of variables controlled by Morelli et al. are included, such as predator size and speed, territorial or non-territorial species, distance to a refuge, morphological defense strategies, and being in a generally good condition. While Morelli et al. focused on different bird species to find phylogenetic signals in FID values, Stankowitch and Blumstein highlight available empirical data showing that also in other species, it is possible to observe changes in FID due to group size. Morphological defense strategies in species different from birds (e.g., lizards or insects) might be an influential aspect of why FID sometimes decreases with group size. Such a decrease is also observed in some fish whose shoaling behavior might be the reason for shorter FID (see ibid., p. 2629). This means that claims about FID and its determinants should include the specification of species to which they are applicable.

The meta-analysis of Stankowitch and Blumstein is based on 116 publications investigating the effect of different factors on FID. The effects of certain variables on FID, as presented by the individual publications out of the 116 in total, were weighted in the meta-analysis by applying the Schmidt-Hunter method. This means that a publication's sample size (of observed populations) is weighted in relation to the total sample size that is included by all publications involved in the meta-analysis (see ibid., p. 2628). The results of this meta-analysis show that, in some species, like certain fish, FID decreases with group size while it increases in others, like water birds (see ibid., p. 2631). The overall network of variables responsible for affecting FID is visualized as follows (Fig. 1):

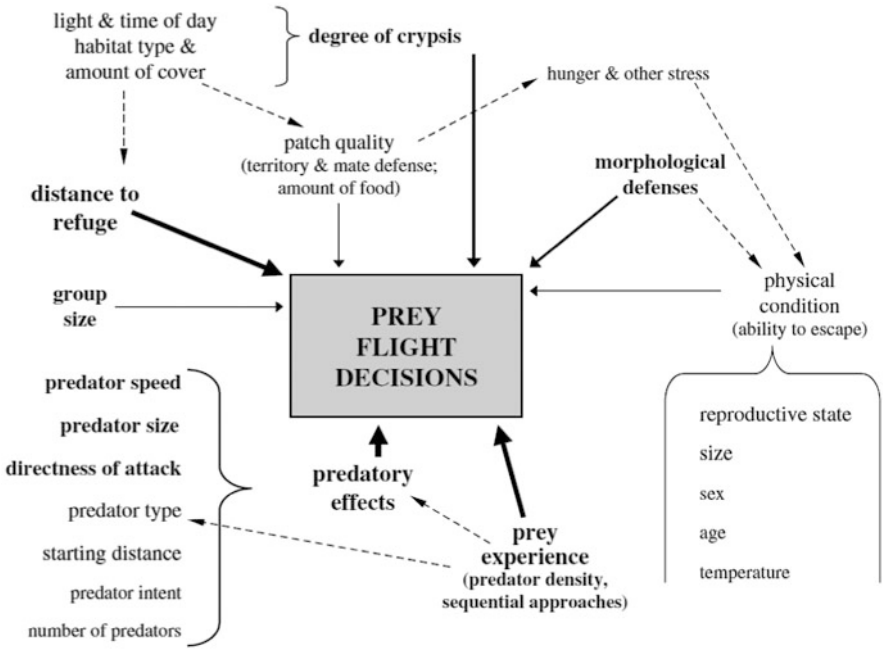


Fig. 1 A visual summary by Stankowich and Blumstein showing potential factors of influence for flight initiation distance in animals. The size of solid lines indicates relative strength of influence. Dotted lines indicate possible indirect relationships. Figure adapted and reproduced from Stankowich and Blumstein (2005). All the copyrights are attributed to the authors

For enactivists, this network is a well-elaborated instance of decentralized, homogeneous, dynamically interacting entities and processes that constitute risk evaluation in birds, which are autopoietic, adaptive organisms with a developmental history and individual learning chronicles. The cited factors modulate certain states of action-readiness, the disposition to flee in specific situations. Terms such as "information processing," "computation," and "mental representations" are not applied in this context to understand how FID is determined. Although we do not claim that cognitivist interpretations and explanations of FID's changes are impossible, such ethological investigations tend to make, from our perspective, a stronger case for enactivist, non-representationalist understandings of risk evaluation. The focus on exogenous aspects of performances supports externalist positions of cognition research that are available in 4E theories in general and enactivism in particular.

4 Conclusion

Enactivism is confronted with methodological challenges such as the motley crew argument and the lack of commitment to explanatory frameworks. In this paper, we argued that ethology holds premises, concepts, and explanation procedures that are easily linked to enactivist thinking and help answer its methodological challenges. Especially the ethological focus on action-readiness, its constituents, and conditions of change matches well with one of enactivism's core claim that cognition consists in the exercise of (skillful) know-how expressed by embodied action in appropriate situations. In both approaches, behavioral profiles are the subject of research, while ethologists are empirically trained to identify biological and non-biological aspects that determine the spectrum of (inter-)actions an organism can execute. Their empirical training is particularly suited for supporting enactivists to meet the motley crew argument since ethologists identify, describe, and keep track of distributed entities and processes that bring behavioral profiles about. We presented an ethological case study of fleeing behavior in European gregarious birds, which investigated the influence of several factors on their flight initiation distance. The case study shows that birds' risk evaluation and fleeing behavior change due to their flock size. It is hence most likely that risk evaluation and fleeing behavior are socially constituted in gregarious birds as the performances of their flock behaviorally enslave them. A further discussed meta-analysis of studies on animals' fleeing behavior corroborates the findings. We claim that the ethological case study and the ethological meta-analysis of similar cases involve explanation procedures enactivists should accept. By doing so, they not only commit to a specific explanatory regime for their research but also answer the motley crew argument by supporting research practices that spot constitutive networks of cognitive performances.

References

Abramova K, Slors M (2019) Mechanistic explanation for enactive sociality. Phenomenol Cogn Sci 18:1–24
Allen C (1998) Assessing animal cognition: ethological and philosophical perspectives. J Anim Sci 76:42–47
Avila-Garcia O, Cañamero L (2005) Hormonal modulation of perception in motivation-based action selection architectures. In: Proceedings of the Symposium on Agents that Want and Like
Barandiaran XE (2017) Autonomy and enactivism: towards a theory of sensorimotor autonomous agency. Topoi 36:409–430
Bateson P, Laland KN (2013) Tinbergen's four questions: an appreciation and an update. Trends Ecol Evol 28:712–718
Bekoff M (1999) Cognitive ethology. In: Bechtel W, Graham G (eds) A companion to cognitive science. Blackwell Publisher, Oxford
Blumstein DT (2006) Developing an evolutionary ecology of fear: how life history and natural history traits affect disturbance tolerance in birds. Anim Behav 71:389–399
Bower M, Gallagher S (2013) Bodily affectivity: Prenoetic elements in enactive perception. Phenomenol Mind 2:108–131
Burger J, Gochfeld M, Jenkins CD, Lesser F (2010) Effect of approaching boats on nesting black skimmers: using response distances to establish protective buffer zones. J Wildl Manag 74:102–108
Burghardt G (1973) Instinct and innate behavior: toward an ethological psychology. In: Nevin J, Reynolds G (eds) The study of behavior. Scott Foresman, Glenview, pp 322–400
Carew TJ, Hawkins RD, Kandel ER (1983) Differential classical conditioning of a defensive withdrawal reflex in Aplysia californica. Science 219(4583):397–400
Casper M-O (2019) Social enactivism: on situating high-level cognitive states and processes. De Gruyter, New York
Casper M-O, Artese GF (2020) Maintaining coherence in the situated cognition debate: what computationalism cannot offer to a future post-cognitivist science. Adapt Behav 30(1):3–17
Chemero A (2013) Radical embodied cognitive science. Rev Gen Psychol 17(2):145–150
Colombetti G, Zavala E (2019) Are emotional states based in the brain? A critique of affective brainocentrism from a physiological perspective. Biolo Philos 34:Article 45
Cosmelli D, Thompson E (2010) Embodiment or envatment? Reflections on the bodily basis of consciousness. In: Stewart J, Gapenne O, Di Paolo E (eds) Enaction: towards a new paradigm of cognitive science. MIT Press, Cambridge, MA
Craver C, Darden L (2013) In search of mechanisms. In: Discoveries across the life sciences. Chicago University Press, Chicago
Darwin CR (1881) The formation of vegetable mould through the action of earthworms. John Murray, London
Dewey J (1896) The reflex arc concept in psychology. Psychol Rev 3:357–370
Di Paolo EA (2005) Autopoiesis, adaptivity, teleology, agency. Phenomenol Cogn Sci 4:97–125
Di Paolo EA (2009) Extended life. Topoi 28:9–21
Di Paolo EA (2018) The enactive conception of life. In: Newen A, Gallagher S, de Bruin L (eds) The Oxford handbook of cognition: embodied, embedded, enactive and extended. Oxford University Press, New York, pp 71–94
Di Paolo EA, Buhrmann T, Barandiaran XE (2017) Sensorimotor life: an enactive proposal. Oxford University Press, New York
Dotov D (2014) Putting reins on the brain. How the body and environment use it. Front Hum Neurosci 8:795
Eliasmith C (2009) Dynamics, control and cognition. In: Robbins P, Aydede M (eds) The Cambridge handbook of situated cognition. Cambridge University Press, Cambridge, pp 134–154
Favela LH (2020) Dynamical systems theory in cognitive science and neuroscience. Philos Compass 15(8):e12695. https://doi.org/10.1111/phc3.12695

Fernández-Juricic E, Jimenez MD, Lucas E (2002) Factors affecting intra- and inter-specific variations in the difference between alert distance and flight distance for birds in forested habitats. Can J Zool 80:1212–1220
Freeman WJ (2000) How brains make up their minds. Columbia University Press, New York
Frijda NH (2007) The laws of the emotions. Lawrence Erlbaum Associates, Mahwah, NJ
Fuchs T (2018) The cyclical time of the body and its relation to linear time. J Conscious Stud 25:47–65
Gallagher S (2005) How the body shapes the mind. Oxford University Press, New York
Gallagher S (2009) Philosophical antecedents to situated cognition. In: Robbins P, Aydede M (eds) Cambridge handbook of situated cognition. Cambridge University Press, Cambridge
Gallagher S (2017) Enactivist interventions: rethinking the mind. Oxford University Press, New York
Gallagher S (2018) New mechanisms and the enactivist concept of constitution. In: Guta MP (ed) The metaphysics of consciousness. Routledge, Oxfordshire, pp 207–220
Garson J (2011) Selected effects and causal role functions in the brain: the case for an etiological approach to neuroscience. Biol Philos 26:547–565
Glover HK, Weston MA, Maguire GS, Miller KK, Christie BA (2011) Towards ecologically meaningful and socially acceptable buffers: response distances of shorebirds in Victoria, Australia, to human disturbance. Landsc Urban Plan 103:326–334
Hutchins E (2010) Cognitive ecology. Topics in cognitive. Science 2:705–715
Hutto DD (2005) Knowing what? Radical versus conservative enactivism. Phenomenol Cogn Sci 4(4):389–405
Immelmann K (1983) Introduction to ethology. Plenum Press, New York
Jamieson D, Bekoff M (1992) On aims and methods of cognitive ethology. In: Proceedings of the Biennial Meeting of the Philosophy of Science Association Vol. 1992, Vol 2 (Symposia and invited papers). The University of Chicago Press, Chicago, pp 110–124
Kaplan A (1962/2017) The conduct of inquiry. Methodology for behavioral science. Routledge, New York
Käufer S, Chemero A (2021) Phenomenology: an introduction, 2nd edn. Polity, London
Kirchhoff MD, Kiverstein J (2020) Attuning to the world: the diachronic constitution of the extended conscious mind. Front Psychol. https://doi.org/10.3389/fpsyg.2020.01966
Kiverstein J, Clark A (2009) Introduction: mind embodied, embedded, enacted: one church or many? Topoi 28(1):1–7
Kiverstein J, Miller M (2015) The embodied brain: towards a radical embodied cognitive neuroscience. Front Hum Neurosci 9:Article 237
Kiverstein JD, Rietveld E (2018) Reconceiving representation-hungry cognition: an ecological-enactive proposal. Adapt Behav 26:147–163
Lamb M, Chemero A (2014) Structure and application of dynamical models in cognitive science. Proc Annu Meeting Cogn Sci Soc 36:809–814
Lehner PN (1979) Handbook of ethological methods. Garland STPM Press, New York
Lima SL (1990) Protective cover and the use of space: different strategies in finches. Oikos 58:151–158
Lima SL, Dill LM (1990) Behavioral decisions made under the risk of predation: a review and prospectus. Can J Zool 68:619–640
Mettke-Hofmann C (2014) Cognitive ecology: ecological factors, life-styles, and cognition. Wiley Interdiscip Rev Cogn Sci 5:345–360
Millikan R (1984) Language, thought and other biological categories. MIT Press, Cambridge, MA
Morand-Ferron J, Cole EF, Quinn JL (2016) Studying the evolutionary ecology of cognition in the wild: a review of practical and conceptual changes. Biol Rev 91:367–389
Morelli F, Benedetti Y, Diaz M et al (2019) Contagious fear: escape behavior increases with flock size in European gregarious birds. Ecol Evol 9:6096–6104
Noë A (2021) The enactive approach: a briefer statement, with some remarks on “radical enactivism”. Phenomenol Cogn Sci 20:957–970

O'Regan JK, Noë A (2001) A sensorimotor account of vision and visual consciousness. Behav Brain Sci 24:939–1011
Port RF, van Gelder T (eds) (1995) Mind as motion: explorations in the dynamics of cognition. The MIT Press, Cambridge, MA
Rietveld E, Kiverstein J (2014) A rich landscape of affordances. Ecol Psychol 26:325–352
Samia SM, Blumstein DT, Díaz M et al (2017) Rural-urban differences in escape behavior of European birds across a latitudinal gradient. Front Ecol Evol 5:Article 66
Satne G (2015) The social roots of normativity. Phenomenol Cogn Sci 14:673–682
Shapiro L (2011) Embodied cognition. Routledge, New York
Silberstein M, Chemero A (2013) Constraints on localization and decomposition as explanatory strategies in the biological sciences. Philos Sci 80:958–970
Stankowich T, Blumstein DT (2005) Fear in animals: a meta-analysis and review of risk assessment. Proc R Soc B 272:2627–2634
Stepp N, Chemero A, Turvey MT (2011) Philosophy for the rest of cognitive science. Top Cogn Sci 3:425–437
Thompson E (2007) Mind in life: biology, phenomenology, and the sciences of the mind. Harvard University Press, Cambridge, MA
Thompson E (2018) Review of Daniel D. Hutto and Erik Myin, evolving enactivism: basic minds meet content. *Notre Dame Philosophical Reviews*. Retrieved from https://ndpr.nd.edu/reviews/evolving-enactivism-basic-minds-meet-content/
Thompson E, Cosmelli D (2011) Brain in a vat or body in a world? Brainbound versus enactive views of experience. Philos Top 39:163–180
Tinbergen N (1963) On aims and methods of ethology. Z Tierpsychol 20:410–433
Toates FM (1986) Motivational systems. Cambridge University Press, Cambridge, NY
Van Orden G, Hollis G, Wallot S (2012) The blue-collar brain. Front Physiol 3:207
Varela F (1991) Organism: a meshwork of selfless selves. In: Tauber AI (ed) Organism and the origin of self. Kluwer Academic Publishers, Dordrecht
Varela F (1997) Patterns of life: intertwining identity and cognition. Brain Cogn 34:72–87
Varela F (1999) Ethical know-how: action, wisdom, and cognition. Stanford University Press, Stanford, CA
Varela F, Maturana HR, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. Biosystems 5:n187
Varela T, Thompson E, Rosch E (1991/2017) The embodied mind. The MIT Press, Cambridge/London
Vogel DHV, Jording M, Kupke C, Vogeley K (2020) The temporality of situated cognition. Front Psychol 11:546212
Ward D, Silverman D, Villalobos M (2017) Introduction: the varieties of enactivism. Topoi 36:365–375
Weiskopf DA (2011) Models and mechanisms in psychological explanation. Synthese 183:Article 313
Wheeler M (2010) Extended functionalism. In: Menary R (ed) The extended mind. MIT Press, Cambridge
Winkel G, Saegert S, Evans GW (2009) An ecological perspective on theory, methods, and analysis in environmental psychology: advances and challenges. J Environ Psychol 29:318–328
Yasué M (2005) The effects of human presence, flock size and prey density on shorebird foraging rates. J Ethol 23:199–204

Causal Closure, Synaptic Transmission and Emergent Mental Properties

Giacomo Zanotti

Abstract The causal argument for physicalism about the mind has received a lot of attention. In particular, the literature has focused on the main premise of the argument, namely the causal closure principle (CC). In this article, I present and discuss the so-called argument from physiology, that is widely regarded as the most convincing line of reasoning in favour of CC. When it comes to providing empirical grounds for the argument from physiology, the most promising move the physicalist can opt for is to focus on the mechanisms of synaptic transmission. Here, I argue that the argument from physiology can provide support for CC only if evidence concerning synaptic transmission is combined with a non-innocent assumption about the internal causal organisation of the nervous system. I contend that this assumption should be vindicated. Unfortunately, this does not seem to be possible at the moment.

Keywords Causal closure principle · Physicalism · Emergent mental properties · Argument from physiology · Mental causation

1 The Causal Argument for Physicalism

Physicalism about the mind is arguably the prevailing position in the contemporary debate on the nature of mental states. Yet, it is not a unitary one. First, a distinction must be made between identity theories and physicalist views that take mental properties to be nothing over and above physical properties without positing (type) identities. Most notably, however, things get complicated when it comes to providing a more precise characterisation of nothing-over-and-aboveness. For this purpose, advocates of physicalism have resorted to different metaphysical relations, ranging from realisation (Melnyk 2006) to constitution (Pereboom 2011) and grounding (Dasgupta 2014) – this list is not meant to be exhaustive – but the question is still

G. Zanotti (✉)
Politecnico di Milano, Milano, Italy
e-mail: giacomo.zanotti@polimi.it

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_4

open. True, some common denominators can be found. In particular, starting with Lewis (1983), it has been argued that all physicalists share at least the commitment to the claim that mental properties are metaphysically supervenient on physical properties.[1] That said, a consensus on a precise definition of physicalism is still lacking.

Dualism is the traditional alternative to physicalism. In this article, I will focus on *property* dualism, and more specifically on its emergence-based formulations. According to the advocates of emergentist dualism, mental properties are *strongly* emergent from physical properties: they are fundamental, only nomologically – as opposed to metaphysically – necessitated by physical properties.[2] In addition, strongly emergent mental properties are often taken to have fundamentally novel causal powers, not possessed by their physical emergence bases (McLaughlin 1992; O'Connor 1994; Kim 1999).

In what follows, I will often refer to the contraposition between physicalism and dualism in terms of weak versus strong emergence.[3] Weakly emergent mental properties are acceptable from a physicalist perspective: they are non-fundamental properties, metaphysically necessitated by physical properties, and lacking novel causal powers. Notice that I am admittedly oversimplifying. A detailed discussion of physicalism and dualism would require a long detour I cannot afford. However, the characterisations I have provided should allow us to address the main subject of this article, that is the causal argument for physicalism.

Among the arguments for physicalism, the causal one has probably received the most attention in the literature.[4] The line of reasoning hinges on the incompatibility between a dualist view and the following premises:

[1] This has been disputed (Montero 2013; Montero and Brown 2018; see Alter 2021 for a reply). However, the greatest majority of authors take metaphysical supervenience to be the physicalist's minimal commitment.

[2] Wilson (2015, 2021) calls into question this way of conceiving strong emergence, arguing that there could be cases in which properties that we would tend to regard as strongly emergent are metaphysically necessitated by physical properties – e.g. a Malebranchean scenario in which God always causes the instantiation of fundamental mental properties upon the occasion of physical properties. However, for our purpose, a modality-based definition of strong emergence will do the work (among others, see Chalmers 2006; Noordhof 2010).

[3] To be clear, these are not the only possible solutions to the mind-body problem. In particular, panpsychism has been receiving a lot of attention in the recent literature (see Bruntrup and Jaskolla 2016; Goff 2017). However, the debate seems to be still largely driven by the dichotomy between physicalism and dualism. This is also the picture emerging from the latest PhilPapers survey (https://survey2020.philpeople.org), in which 51.9% of the participants leaned towards physicalism, 32.1% towards anti-physicalism, and 15.9% opted for 'other'. A more specific question on consciousness in the survey reveals that 22% of the participants leaned towards dualism, while only 7% were sympathetic to panpsychism.

[4] I focus on the line of reasoning that is discussed in the contemporary debate (see Papineau 2001, 2002). However, it is worth highlighting that the causal argument has some antecedents in the history of Western philosophy. It is sufficient to think about Princess Elisabeth of Bohemia writing to Descartes that it would be easier 'to concede matter and extension to the soul than to concede the capacity to move a body and to be moved by it to an immaterial thing' (III 685 AT, in Shapiro 2007,

(1) mental properties have physical effects;
(2) all physical effects are fully caused by purely *physical* prior histories (the causal closure principle);[5]
(3) the physical effects of mental properties are not systematically overdetermined by physical properties.[6]

If all three of these claims are true, then dualists are in serious trouble. On the one hand, they hold that *sui generis* mental properties, metaphysically distinct from the physical properties they emerge from, are among the causes of our behaviours. On the other hand, (2) implies that our behaviours already have a history of *sufficient* physical causes. The only way to ease this tension would be to argue that our behaviour is systematically overdetermined by mental and physical causes. However, this possibility is precluded by (3).

To facilitate the discussion, let us consider a simplified case of mental causation in which a mental property M and a physical property P compete in the production of a physical event *e*.[7] Four options seem to be available to the dualist when presented with the causal argument:

A. First, the dualist could reject (1). This way, M would turn out to be merely epiphenomenal. No overdetermination would be involved since M would have no causal power.
B. Otherwise, the dualist could reject the causal closure principle. Clearly, if there is no constraint on the nature of the events that can be part of the sufficient causal history of *e*, then there is nothing wrong with M playing an *ineliminable* role in the production of *e*.[8] Again, no overdetermination would be involved, since no physical property P would compete with M.
C. A third possibility is to deny (3). Both M and P would count as individually sufficient causes of *e*, that would turn out to be genuinely overdetermined. Note that in this case, unlike in (B), *e* would occur even if M – or P – failed to be

p. 68). Interestingly, the same line of reasoning can be found in Lucretius' *De rerum natura*, III 162–168.

[5] This version of the principle is employed in Papineau (2002). I take a closer look at the possible formulations of the principle in the next section.

[6] According to the standard definition of overdetermination, *A* and *B* overdetermine an event *e* iff:

(i) *A* and *B* are distinct events;
(ii) *A* is sufficient for causing *e;*
(iii) *B* is sufficient for causing *e;*
(iv) If *A* did not occur, *e* would still occur;
(v) If *B* did not occur, *e* would still occur.

[7] In the debate on causal closure, the standard notion of event is Kim's (1976) one, according to which events are 'property exemplifications' that can be represented as ordered triples $<x, P, t>$. Focusing on monadic events, an event consists of an object *x* instantiating a property *P* at a moment *t*.

[8] By ineliminable, I mean that if M failed to be instantiated, then *e* would not occur. No other property would be instantiated in place of M, making up for the absence of M's causal powers.

instantiated. The instantiation of the remaining property would be sufficient for the instantiation of *e*.

D. Lastly, the dualist could reject property dualism and admit that the relevant causal powers of M simply coincide with the relevant causal powers of P – the easiest way to do it is probably to admit that M and P are at least token-identical. In this case, M could still be taken to be the cause of *e* without violating (2), given that mental causation would turn out to be just physical causation.

Although there are four possible ways out from the inconsistency, the physicalist takes for granted the truth of premises (2) and (3) of the causal argument, significantly reducing the range of moves the dualist can opt for. As a result, the conclusion of the argument comes in the form of a dilemma: either mental properties are merely epiphenomenal, or dualism is false. Needless to say, neither option is desirable for the dualist.

In discussing the causal argument, I will not consider the possibility of claiming that mental and physical properties overdetermine their effects in all cases in which a given physical event is supposed to be caused by a conscious occurrence. True, it has been argued that systematic overdetermination in cases of mental causation is not particularly problematic.[9] However, in case M and P overdetermined their effect *e*, *e* would occur even if M failed to be instantiated. On the contrary, the dualist I have in mind holds that mental properties play an *ineliminable* role in the production of our behaviour. Arguing that *sui generis* mental properties have causal powers that systematically overdetermine their effects seems to be a metaphysically onerous, ad hoc move. For these reasons, I will just assume (3) along with the physicalist. My aim is rather to show that the rejection of (2) is a viable option for the dualist.

I start by outlining the two competing models of mental causation the physicalist and the dualist appeal to (Sect. 2). After that, I focus on the causal closure principle. After arguing for a specific formulation of the principle (Sect. 3), I consider the arguments that have been provided in its favour (Sect. 4). In particular, I focus on Papineau's argument 'from physiology'.[10] I argue that the evidence from neurophysiology Papineau has in mind does not provide direct reasons in favour of the causal closure principle. A further assumption concerning the internal causal organisation of our nervous system is needed. However, such an assumption may be harder to justify than the causal closure principle itself (Sect. 5).

Before proceeding, however, let me say something about the rejection of (1). The dualist's first reaction to the causal argument could be to bite the bullet and accept

[9] See Sider (2003); a different view is defended in Bernstein (2016). The literature on overdetermination is vast and complex. In addition, what is usually at stake in the debate is the possibility of non-reductive, physicalist models of overdetermining mental causation. As far as I can see, arguing for systematic overdetermination within a dualist framework is way more difficult.

[10] Note that this is not the only argument we have available. In particular, the argument 'from fundamental forces' (Papineau 2001) is frequently discussed in the literature. However, the argument from physiology is 'broadly considered much more convincing' (Dimitrijević 2020), and it seems to be the one even Papineau insists upon in his latest contributions (see Papineau 2020).

that *sui generis* mental properties have no causal powers upon the physical domain. This view, which falls under the label of epiphenomenalism, has an illustrious history. Among others, Malebranche's occasionalism and Leibniz's pre-established harmony can be regarded as theoretical prototypes of epiphenomenalism. Both reject the possibility of any causal interaction between the mind – or, more precisely, the soul – and the body. More recently, epiphenomenalism has been defended by philosophers such as Campbell (1970), Jackson (1982), and Robinson (2019).[11] Chalmers (1996) expressed a certain sympathy for non-interactionist forms of dualism as well, although he is more cautious in his later works (see Chalmers 2010).

Interestingly, it is sometimes suggested that there is some empirical evidence in favour of epiphenomenalism. More specifically, reference is made to Libet's (1985) famous experiments. To make a long story short, these experiments would show that some neural activations, that would be responsible for the initiation of simple movements, significantly precede the conscious decision to perform those movements. Therefore, physical properties would pre-empt mental properties of their causal role. As far as I can see, there are good reasons for being suspicious. First, it should be kept in mind that some methodological aspects of these studies have been disputed (Gomes 2002, Pockett and Purdy 2011; see also Lavazza 2016). Furthermore, the philosophical implications of the obtained results are far from being clear (Mele 2014; Baumeister et al. 2018).

In addition to this, some arguments against epiphenomenalism have been provided in the literature. Among others, the one from natural selection and the self-stultification objection are worth recalling. In broad brushstrokes, the former is a line of reasoning leveraging the intuition that, if consciousness were epiphenomenal, then its evolution would be inexplicable (Popper and Eccles 1977).[12] The latter, instead, is an argument to the effect that epiphenomenalism would be simply incompatible with our knowledge of our own mental states (see De Brigard 2014 for a discussion). True, advocates of epiphenomenalism could easily challenge the evolutionary argument. In particular, they could argue that the evolution of consciousness is 'a sort of byproduct' of physical evolution (Chalmers 2010, p. 131). Resisting the self-stultification objection, however, may be more complex.

For space reasons, I cannot afford to go into details. Let me just add that, even if compelling counter-arguments to the self-stultification objection were provided, giving up the causal efficacy of consciousness would not be so easy. Epiphenomenalism is at odds with our basic intuitions about the way our behaviour is influenced by our psychological life. Here is how Fodor (1989, p. 77) puts it:

> If it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying…, if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.

[11] See also Baysan (2020).

[12] See also James (1879) on the evolutionary utility of pleasure and pain.

Note that I am not assuming that our intuitions are always infallible in the context of theory choice in metaphysics, nor that they are in this specific case. All I am pointing out is that the implications of epiphenomenalism are extremely counterintuitive. In what follows, I will assume that the dualist has good reasons for avoiding the epiphenomenalistic apocalypse. In fact, I take the causal efficacy of mental states to be a *desideratum*, regardless of the metaphysical theory one ends up adopting.

With these premises, we can proceed with our analysis. In the next section, I will briefly present and discuss the two different models of mental causation the physicalist and the dualist are committed to.

2 Two Models of Mental Causation

As we have seen, four claims are at stake in the causal argument. I have assumed that both the physicalist and the dualist agree on two points:

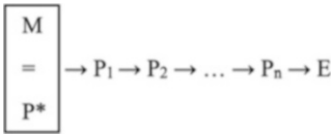
- causal efficacy: mental properties have physical effects;
- no overdetermination: the physical effects of mental properties are not overdetermined by physical properties.

What they disagree upon are the following claims:

- causal closure: all physical effects are fully caused by purely physical prior histories;
- property dualism: mental properties are something over and above physical properties.

To make things easier, let us consider a simple scenario in which a subject *S* is thirsty, and their experience of thirst (M) seems to be the cause of a chain of physical events (P_1, P_2, … P_n) – motor neurons firing, sarcomeres contracting, and so on – that ultimately result in *S* ingesting water (E). Include another physical state P*, that co-varies with M and can be referred to as M's emergence base. What is at stake in the argument is the precise nature of the relation between M and P*.

The physicalist avoids the inconsistency among the premises of the causal argument by rejecting (property dualism). Here is a schematic illustration of the physicalist model of mental causation – the arrows represent causal relations, going from causes to effects:

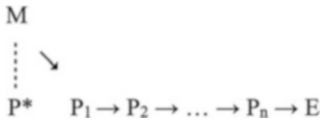


According to the physicalist, M and P* are at least token-identical, and the distinction that is made is purely conceptual.[13] As a result, there are no two sources of causal power.

On the contrary, the dualist – or at least, the dualist I have in mind – aims at maintaining both (property dualism) and (causal efficacy). In their view, M and P* are still instantiated together. However, they are metaphysically distinct properties. M strongly emerges from P*, and the co-variation between them is accounted for in terms of contingent psychophysical laws. In addition, there should be room for a causal arrow that goes from M to P_1. Clearly, the plausibility of this model depends on the possibility of rejecting the causal closure principle.

At this stage, the physicalist could react by arguing that this project is utterly unrealistic, even conceding that (causal closure) may not hold. It could be argued that if both M and P* alone were sufficient for P_1 and were instantiated at the same time, (no overdetermination) would be violated. What is more, even if overdetermination worries are mitigated – e.g. by adopting a non-oomphy notion of causation *à la* Lewis/Woodward – independent considerations make this model implausible. If we admit that M and P* co-occur at *t* and that P* alone is enough for bringing about P_1, why should we posit another sufficient cause? There is a sense in which assigning a causal role to M appears to be a metaphysically onerous, dispensable move.

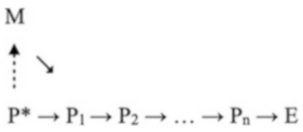
The dualist can easily address the issue by pointing out that, with (causal closure) out of the picture, nothing forces us to presuppose that the exercise of P*'s causal power is sufficient for causing P_1, or even that P* has a causal role at all. Once (causal closure) is rejected, dualists have two options. First, they can argue that M is the only cause of P_1:



Arguably, this sounds like heresy to the physicalist. However, once (causal closure) is dismissed, there is nothing absurd in the hypothesis of a mixed causal chain, that presumably starts before M with a series of physical – and possibly mental – events, involves M, and finally ends with E.

[13] Whether non-reductive versions of physicalism are ruled out by the causal argument is a controversial issue that has been extensively debated. In particular, Kim's (1998, 2005) exclusion argument against non-reductive physicalism is worth mentioning. I will not go into the details of Kim's line of reasoning, for which a number of non-reductive solutions have been suggested (among others, see Bennett 2003). In what follows, I will concede that the physicalist can reject dualism on the grounds of the causal argument without being committed to a strong identity thesis between mental and physical properties.

Alternatively, the dualist can argue that M and P* are co-causes of P_1. That is, the causal powers of both M and P* are required to bring about P_1. In particular, I have in mind the model defended in Lowe (2000, 2003):



In Lowe's view, P* is the physical base for the emergence of M. More precisely, the instantiation of P* is the *cause* of the instantiation of M at *t*. Crucially, both M and P* count as causes of P_1 at *t*. However, since they are taken to be co-causes, they do not overdetermine P_1. Neither P* nor M can bring about P_1 alone, and P_1 would not occur in case M or P failed to be instantiated at *t*.

In the rest of this article, I will focus on the contraposition between physicalism and emergentist dualism *à la* Lowe, that seems to me to be the most promising interactionist model of mental causation the dualist can resort to. In both cases, P*, that is the physical property serving as emergence base for M, plays a causal role with respect to P_1. What is at issue, besides the precise nature of the emergence relation between M and P*, is whether P*'s causal powers are sufficient for the occurrence of P_1 or the contribution of a *sui generis* mental property is also needed.

3 The Causal Closure Principle

At this stage, we are in the position to carefully consider the causal closure principle, starting with its formulation. As pointed out in the literature, one of the major difficulties the physicalist must deal with is to provide an *adequately strong* version of the principle (Lowe 2000).[14] If the adopted formulation is too weak, the causal argument fails to provide reasons for believing that physicalism is true – or, more precisely, that interactionist dualism is false. If the principle is too strong, on the other hand, it turns out to be almost indistinguishable from the conclusion of the causal argument. In addition, intuitively, the stronger is the principle, the harder it is to vindicate.

When providing an outline of the causal argument, I referred to the causal closure principle as the claim that 'all physical effects are fully caused by purely *physical* prior histories'. This formulation, however, is to a large extent ambiguous. In particular, it is not entirely clear what 'fully' means in this context. On the one hand, one could interpret the principle as stating that, at every moment *t* in the causal history of a given physical effect *e*, if *e* has a cause at *t*, then *e* has a sufficient physical cause at *t*. On the other hand, one could take the principle to state that the

[14] For an overview of the provided formulations, see Gibb (2015).

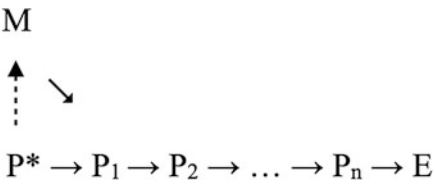
causal histories of physical events are *exclusively* made up of physical events. Let us take a closer look at these two different readings.

The first interpretation is arguably the most common in the debate. I will refer to it as the *weak* formulation of the causal closure principle (WCC):

(WCC) if a physical event *e* has a cause at *t*, it has a sufficient physical cause at *t*.[15]

The first thing to be noticed is that WCC is stronger than another similar principle that can be found in the literature, according to which 'if a physical event has a cause that occurs at *t*, it has a physical cause that occurs at *t*' (Kim 2005, p. 43). If the sufficiency requirement is not specified, then the principle is way too permissive with respect to the acceptability of causally efficacious *sui generis* mental properties. Most notably, it is compatible with the possibility that both physical and non-physical events are needed at a moment *t* to bring about a given physical effect. Consider the dualist scenario in which M – a *sui generis* mental property – is instantiated at *t* together with its emergence base P*. Suppose that both M and P* are required for the production of a given physical effect *e*. In this model, even if M plays an ineliminable causal role in the production of *e*, *e* has a physical cause at *t*, so the causal closure principle without the sufficiency requirement is respected. Clearly, this is not enough for the physicalist, who aims at ruling out *sui generis* mental causation.

WCC, on the contrary, is explicit about the fact that *e* has a *sufficient* physical cause at *t*. Prima facie, if M and P* were co-causes of *e* in the way just described, P* would not count as a sufficient cause of *e* at *t,* and WCC would be violated. Apparently, if WCC holds, the only way for M to somehow cause *e* would be to be an *overdetermining* cause of it. The problem is that, on closer inspection, WCC too turns out to be compatible with a dualist, interactionist view of mental causation. Let us consider Lowe's emergentist model of mental causation I have outlined in the last section. As we have seen, in Lowe's view, the instantiation of a physical property P* at *t* simultaneously causes the emergence of a metaphysically distinct mental property M. Importantly, M and P* are co-causes of P_1 at *t*:



[15] Specifying *t* is needed to rule out the possibility that a physical event *e* occurring at t_2 has a purely mental cause M at t_1 that is in its turn the effect of a physical cause P at *t*. If the causal closure principle simply stated that 'all physical effects have sufficient physical causes' (as in Papineau 1998), *e* could have a purely mental cause at t_1 and still respect the principle. Since causation is (usually regarded as) a transitive relation, the fact that P causes M at *t* would be enough for granting that *e* has a sufficient physical cause in the scenario just described.

In this case, even if M plays an ineliminable role in the production of P_1, P* transitively counts as a sufficient cause of P_1 at *t*. P* is sufficient for causing M at *t*, that together with P* is sufficient for causing P_1. Unless we reject the possibility of simultaneous causation, there seems to be room for an interactionist dualism that respects WCC.

This is something the physicalist is not willing to accept. While it is hardly questionable that Lowe's model is compatible with WCC, it is clear that it fails to respect the *spirit* of the causal closure principle. What the physicalist really has in mind when resorting to the causal closure principle is arguably something along these lines:

> Pick any physical event [. . .] and trace its causal ancestry or posterity as far as you would like; the principle of causal closure of the physical domain says that this will never take you outside the physical domain. (Kim 1998, p. 40)

This directly leads us to the second interpretation of Papineau's (2002) formulation, that is the *strong* causal closure principle (SCC):

> (SCC) physical events can only have physical causes

Clearly, this prevents the dualist from opting for interactionist models of mental causation *à la* Lowe. Since M is not a physical property, there is no room for the exercise of its causal powers in the causal history of P_1, full stop. Admittedly, SCC may sound appealing. However, the physicalist should be cautious in adopting it in the context of the causal argument for physicalism. If the formulation of the principle is too strong, the causal argument begs the question. In fact, it is not difficult to see how this can happen.

One of the most insidious issues the physicalist must deal with when providing a formulation of the causal closure principle concerns the meaning of 'physical' (Crane and Mellor 1990). In order not to fall prey to Hempel's dilemma, physicalists often opt for a negative definition, according to which 'physical' should be interpreted as 'non-mental'.[16] Hence, SCC turns out to be equivalent to the following principle:

> (SCC*) physical events can only have non-mental causes

SCC*, in its turn, can be rewritten as the claim that *sui generis* mental properties cannot cause physical events. This, however, is suspiciously close to the conclusion the physicalist aims at reaching by means of the causal argument. If not question-begging, the causal argument would turn out to be redundant. In particular, once the possibility of all kinds of mental-to-physical causation is ruled out, it is not clear what role the no-overdetermination premise should have.

I take WCC and SCC to be examples of how different formulations of the causal closure principle can be too weak or too strong. In what follows, I will adopt the following one:

[16] In the sense of not fundamentally mental. Most notably, this view has been defended by Montero and Papineau (2005).

(CC) if a physical event *e* has a sufficient cause, it has an immediate sufficient physical cause[17] (Papineau 2009)

On the one hand, CC is sufficiently strong. The immediacy requirement is meant to rule out the possibility of models *à la* Lowe, in which M is a sort of causal intermediary between P* and P_1, and P* (partly) causes P_1 indirectly by simultaneously causing M.[18] For CC to hold, P* must be a sufficient cause of P_1 and *directly* cause P_1, without the intermediate intervention of M. On the other hand, CC is not *too* strong. The possibility of *sui generis* mental properties having a causal role is still open. Most notably, they could be overdetermining causes of the physical effects CC refers to. To reach the desired conclusion, the physicalist must combine CC with the rejection of systematic overdetermination. Therefore, including CC among the premises of the causal argument does not make the line of reasoning a question-begging one.

At this stage, a precise formulation of the causal closure principle has been provided. In the remainder of this article, I will focus on the way the physicalist argues for its truth.

4 The Argument from Physiology

The arguments in favour of CC that are usually discussed in the literature are the argument 'from fundamental forces' and the argument 'from physiology' (Papineau 2001, 2002). In a nutshell, the former line of reasoning is an inductive one that insists on the fact that a number of prima facie special forces turned out to be reducible to a limited set of fundamental, conservative physical forces.[19] The conclusion is that there are no special mental forces that are irreducible to basic physical forces. The argument from physiology, instead, hinges on the fact that despite the impressive progress in recent physiological research, no trace of special mental forces has been found. On the contrary, physical explanations for a number of biological – and more specifically, neural – phenomena have been provided. The conclusion, once again, is that there is no room for special mental forces. Admittedly, both arguments are not conclusive. However, this is not necessarily a problem for the physicalist. After all, there are many beliefs we entertain without having conclusive reasons for doing

[17] Note that this does not make mediated causation per se problematic. Given a physical event e_1 and its alleged physical effect e_2, CC is perfectly compatible with the possibility of an intermediate physical event e^* that is caused by e_1 and causes e_2. What CC rules out is the possibility of *non-physical* causal intermediacy.

[18] See also Garcia (2014) on ontologically proximal and distal causes.

[19] The argument from fundamental forces should not be confused with the one suggested by Dennett (1991), according to which the exercise of special mental forces would violate the principle of conservation of energy. Despite its initial appeal, this line of reasoning seems to be irremediably flawed; see Papineau (2002), Gibb (2010), Tomasetta (2015).

it. Rejecting the arguments in question solely because of the fact that they are not knock-down ones does not seem fair.

In these pages, I will focus on the argument from physiology. This choice is due to the fact that this line of reasoning is generally considered more convincing. Among other things, unlike the argument from fundamental forces, it does not require one to make the question-begging assumption that mental forces – whatever they might be – are not fundamental (Garcia 2014).

To get started, let us look at the formulation of the argument from physiology provided by Papineau (2001, p. 27):

> [. . .] there is no direct evidence for vital or mental forces. Physiological research reveals no phenomena in living bodies that manifest such forces. All organic processes in living bodies seem to be fully accounted for by normal physical forces.

As far as I can see, two distinct – although inevitably interrelated – components of the argument can be isolated. On the one hand, the emphasis is on the fact that we have failed to detect any kind of action in living bodies that could be ascribed to special mental forces operating. On the other hand, scientists have succeeded in providing physical explanations for a number of biological phenomena and processes. Note that, in both cases, the evidence at stake is arguably the one that is provided by research in neurobiology and neuroscience broadly conceived. After all, if there is something that is affected by *sui generis* mental forces, that is arguably the nervous system. In what follows, I will consider the two components of the argument individually.

4.1 The First Component

Let us start with the claim that physiological research has provided no evidence for the action of *sui generis* mental forces. Referring to the latest advancements in biological sciences, Papineau (2001, p. 31) argues:

> [. . .] these developments made it difficult to go on maintaining that special forces operate inside living bodies. If there were such forces, they could be expected to display some manifestation of their presence. But detailed physiological investigation failed to uncover evidence of anything except familiar physical forces.

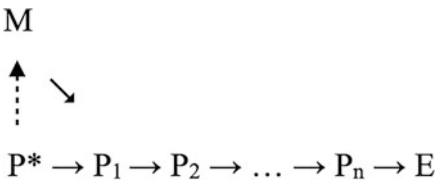
A way to resist this line of reasoning immediately comes to mind. The claim that there are no special mental forces, the dualist may argue, cannot be legitimately inferred from the lack of evidence attesting to such forces' action. The absence of evidence would be regarded as evidence of absence, and this is a typically fallacious move.

I am suspicious about dismissing the line of reasoning in question by simply branding it as a case of *argumentum ad ignorantiam*. If one thinks about it, there seem to be cases in which the absence of evidence can provide reasons for believing that something is actually non-existent. Montero (2003) explicitly addresses this point by discussing the example of ghosts. There is a sense in which the absence of

evidence for ghosts' existence can provide reasons for believing that they actually do not exist. The condition is that we also have knowledge of what really causes noises echoing in the night and the other phenomena that could be traced back to ghosts' action and could have led us to posit their existence in the first place.

Similarly, Montero argues, the absence of evidence for the existence of *sui generis* mental forces can serve as evidence of absence, provided that 'we also have a fairly good understanding of what fundamentally nonmetal force actually causes us to cry out when in pain, and so forth' (2003, p. 185).[20] This sounds pretty reasonable. Hence, it all seems to come down to whether we know enough of the physical processes that are supposed to be the causes of our behaviour. Interestingly, this issue is of the utmost importance also when it comes to assessing the second component of the argument from physiology. For this reason, I will leave the question unanswered for the moment. What I would like to do, now, is to focus on a couple of potential problems that pertain exclusively to the first component.

A crucial assumption of the first component of the argument from physiology is that special mental forces would be empirically detectable. If this were not the case, then appealing to the absence of evidence for their efficacy would be preposterous. The fact that scientists have never observed *sui generis* mental causation is perfectly compatible with the existence of empirically undetectable mental forces. Unfortunately for the physicalist, however, it seems that we cannot exclude the possibility that mental properties' causal contribution is undetectable. Let us consider again Lowe's (2003) model of mental causation:



Lowe insisted on the fact that the exercise of M's causal powers is *invisible*.[21] On the one hand, it is somewhat dubious that the tools that are available to scientists are able to detect causes other than the physical ones. On the other hand, since M and P* are instantiated at the same moment *t*, any external observer is likely to conclude that P* is *immediately* sufficient for the production of P_1. After all, empirical data reveal that the instantiation of P* at *t* is systematically followed by the instantiation of P_1 at t_1. The possibility of a non-physical causal intermediary would not even be considered.[22]

[20] Montero argues that 'while we certainly do not have a complete nonmental account of what we take to be mental causes, we have a good start' (2003, p. 185).

[21] This is ultimately the reason why Montero (2003) is skeptical about the argument from physiology.

[22] Robb (2018) has argued against the invisibility claim. He contends that there is, at least in principle, a way to empirically determine whether strongly emergent mental properties have a

Interestingly, Lowe's model need not be actual to represent an obstacle for the argument from physiology. Even the mere possibility that mental causation is invisible threatens the efficacy of the first component of the argument, since the physicalist cannot exclude that such a possibility actually obtains (Owen 2020). Accepting that mental causation *could* be invisible is enough for undermining the inference from the absence of evidence to the evidence of absence the physicalist relies on.[23]

That said, let us suppose that either Lowe's model is implausible, or that Lowe is wrong about the invisibility of mental causation. Still, there might be some tension when it comes to the claim that 'there is no principled *a priori* reason why 20th-century physiological research should not have uncovered special mental and vital forces' (Montero and Papineau 2016). A potential problem stems from the ambiguous status of the causal closure principle. Here, CC has been presented as a substantial, metaphysical claim about the causal structure of the universe. However, its *methodological* implications should be made clear as well. In fact, as Kim (1996, pp. 147–148) points out, many physicalists 'accept the causal closure of the physical not only as a fundamental metaphysical doctrine but as an indispensable methodological presupposition of the physical sciences'.[24]

The claim that the causal closure principle also has a methodological component seems to be a plausible one. Focusing on physiology and neuroscience, the following methodological precept seems to drive current research:

> (MCC) when accounting for the production of human behaviour, the *explanans* cannot include causes other than physical ones[25]

However, this might have an impact on the physicalist's line of reasoning. With MCC in the picture, the physicalist may be prevented from legitimately resorting to the first component of the argument from physiology. Assume that *sui generis*

---

causal role – more on this in Sect. 5. However, even if Robb is right, mental causation *à la* Lowe is nonetheless *almost* invisible. This is enough for undermining the first component of the argument from physiology.

[23] An anonymous reviewer cast doubts on this point by providing a counterexample: the absence of evidence that there was a tornado an hour ago is good evidence that there was no tornado, even if one cannot exclude the possibility that there was actually a tornado but the damages were immediately and silently repaired. As far as I can see, there is a crucial difference between this scenario and the case of invisible mental causation. We know that tornados can leave evidence in the actual world – and they usually do, which explains why we regard the described scenario as an unlikely one. Provided that we could be wrong, the fact that we have repeatedly observed such evidence allows us to legitimately infer that there was no tornado an hour ago. On the contrary, we do not know whether invisible mental causation takes place in the actual world. This makes the inference from the absence of evidence to the evidence of absence much riskier.

[24] See also Zargar et al. (2020).

[25] Admittedly, MCC is the methodological version of SCC, that is not the principle I am considering. To be consistent, I should take MCC to be the precept that when accounting for the production of behaviour, the *explanans* cannot include causes other than *immediately sufficient* physical ones. As far as I can see, the simplification does not affect in any significant way the considerations made here.

mental properties actually do have a causal role in the production of behaviour. Suppose then that we have a somewhat fine-grained account of what goes on in the brain between, say, the second before a subject's decision to move their left arm and the contraction of their muscles. Even if some gaps were detected in the physical causal chain, neuroscientists would not even think about *sui generis* mental causation. In accordance with MCC, they would simply keep searching for some physical events that may have gone unobserved. *Contra* Montero and Papineau, if scientific investigation in neuroscience is driven by something along the lines of MCC, there seem to be principled reasons to doubt that recent research could have uncovered special mental forces. Even if we had an account of the physical causal processes within the brain that is detailed enough to allow for the detection of causal gaps, such negative evidence about the possibility of non-physical forces operating would simply be mistaken for ignorance about the existence of further physical causes.

At this stage, Montero and Papineau could insist that the causal closure principle has its methodological implications *in virtue of the fact* that it is a strongly empirically supported metaphysical claim. Put another way, scientists would exclusively look for physical causes because *prior* evidence has suggested that physical events have only physical causes. This sounds plausible, although exploring the issue would require a detailed historical analysis. In general, however, it is fair to say that physicalists should be careful when arguing along with Kim that CC has strong methodological implications. More specifically, they have to point out that the metaphysical component is prior – historically and epistemically – to the methodological one. Otherwise, one may suspect that the valid negative evidence for CC is not as much as we tend to think.

4.2 The Second Component

At this stage, let us consider the second component of the argument from physiology, that focuses on the *positive* evidence we have available. In particular, Papineau (2001, p. 31) insists on the development of biology and neuroscience in the twentieth century:

> [. . .] the catalytic role and protein constitution of enzymes were recognized, basic biochemical cycles were identified, and the structure of proteins analyzed, culminating in the discovery of DNA. In the same period, neurophysiological research mapped the body's neuronal network and analysed the electrical mechanisms responsible for neuronal activity.

Briefly, the idea seems to be the following: the more we learn about causal processes in living organisms without finding evidence for *sui generis* mental causation, the less room for *sui generis* mental causation is left. Note that this is not the same as arguing that we have not found any trace of *sui generis* mental causation. What the physicalist insists on, in this case, is that recent research has identified a number of physical causes of our behaviour. The strength of this line of reasoning directly depends on our positive knowledge of the brain's functioning. If we can count on a

sufficiently precise account of the physical processes that are responsible for the production of our behaviour, we may have reasons to think that *sui generis* mental properties do not have a role.

Unfortunately for the physicalist, optimism in the actual explanatory power of neuroscience and physiology may be misplaced. To be clear, I am not adopting an antiscientific stance. On the contrary, I assume that philosophy cannot ignore the empirical evidence we have available. Narrowing it down to the mind, philosophers should not overlook the results that neuroscience and psychobiology have been providing. At the same time, I am not calling into question the import of the advancements that have been made in these disciplines. New tools have opened new possibilities for studying the brain, both structurally and functionally, and our understanding of the physiology and organisation of our nervous system has dramatically improved in recent years. The point is just that unrealistically high expectations seem to be put on the explanatory power of current science.

Coming back to the first component of the argument from physiology, let us consider once again the example of ghosts' existence, that Montero (2003) takes to be a case in which the absence of evidence counts as evidence of absence. As she points out, the sine qua non condition for this to be possible is that we have a 'pretty good understanding of what actually causes those bumps in the night that scare people into thinking their houses are haunted' (2003, p. 185). Accordingly, when we consider the absence of evidence for *sui generis* mental causation and we aim at inferring that there are no special mental forces operating, we need a 'pretty good understanding' of what actually causes our behaviour. Needless to say, it is not enough to know that the activation of specific brain areas has a causal role in the production of certain responses. This coarse-grained knowledge is perfectly compatible with the exercise of non-physical causal powers somewhere in the causal chain. What seems to be needed as background knowledge is a reasonably good approximation of a *complete* and *fine-grained* description of the causal processes that are supposed to produce our behaviour. If this description involves no *sui generis* mental causes, then we may have good reasons to believe in the absence of non-physical causes of behaviour.

A similar level of detail seems to be required in the second component of the argument from physiology. The more psychophysiology and neuroscience are successful in providing detailed, purely physical descriptions of the causal processes that bring about our behaviour, the less *sui generis* mental causes fit in the picture – here, the assumption is that *sui generis* mental causation would have effects on the brain. The fewer links in the causal chain are left unidentified, the more *sui generis* mental causation turns out to be implausible. Again, what the physicalist needs to presuppose to rule out special mental forces seems to be that psychophysiology and neuroscience have already provided a reasonably fine-grained and almost complete account of the way our behaviour is produced by physical processes.

Unfortunately, this confidence may not be justified. Let us focus, along with Owen (2020), on a passage from a recent book by Christof Koch, one of the most prominent neuroscientists of consciousness. Referring to our 'inadequate knowledge

of the prodigious complexity of the brain, from the molecular to the system level', Koch (2019, p. 138) writes:

> The dirty secret of computational neuroscience is that we still do not have a complete dynamic model of the nervous system of the worm *C. elegans*, though it only has 302 nerve cells and its wiring diagram, its connectome, is known. So here we are, trying to understand the human brain, when we do not yet understand the worm brain.[26]

This may seriously undermine the soundness of the physicalist's line of reasoning. On the one hand, the absence of evidence concerning *sui generis* mental causation cannot count as evidence of absence, since we lack that 'pretty good understanding' of the brain's functioning that would make the inference legitimate. On the other hand, since our knowledge of the causal interactions within the brain is far from being complete, we cannot exclude the possibility of non-physical causes intervening at some stage of the causal process.

Taking stock, both the components of the argument from physiology seem to rely on an excessively optimistic assumption concerning our knowledge about the functioning of the nervous system, which is arguably what would be influenced by *sui generis* mental causes. At this stage, physicalists may reply that we need not assume that we have a fine-grained and complete account of the causal processes within the brain. What we need, they may argue, is a reasonably detailed description of the functioning of the nervous system's working units.

In a recent discussion of the causal argument, Papineau (2020, p. 16) explicitly refers to such knowledge as the empirical ground for the argument from physiology:

> It was only in the middle of the twentieth century that a detailed understanding of the electrochemical workings of neurons convinced the scientific mainstream that there is no place for *sui generis* mental forces.

The point is that we know which kinds of neurotransmitters, receptors, and molecules are involved in synaptic transmission, and we can count on a reasonably precise reconstruction of the way they causally interact. Not by chance, the process of synaptic transmission is often used in the philosophy of science as an example of a phenomenon that was accounted for in mechanistic terms (among others, see Craver 2007). Whether this kind of evidence is sufficient to vindicate CC is the question I will address in the remainder of this article.

5 Synaptic Transmission, Causal Closure and Emergence

When it comes to the mechanisms of synaptic transmission, a sufficiently detailed story of the involved causal processes seems to be available. Given two neurons n_1 and n_2, the electrical impulse reaching the synaptic vesicles at the end of n_1's axon arguably counts as the immediate, sufficient physical cause of neurotransmitters

[26] See also Garcia (2014); Di Francesco and Tomasetta (2015).

being released in the synaptic cleft between n_1 and n_2. In its turn, the release of neurotransmitters in the cleft seems to be the immediate, sufficient physical cause of the activation of specific receptors on the post-synaptic membrane of n_2. Ultimately, the result is that n_2 is either excited or inhibited. Regardless of the oversimplification, the metaphysical take-home message is clear. At least prima facie, there seems to be evidence for what I will refer to as the *synaptic* causal closure thesis:

> (Synaptic-CC) physical events *within the synaptic micro-system* have immediate and sufficient physical causes

True, one may argue that we cannot be *sure* that we know everything about the causal steps involved in the mechanisms in question. Hence, there could be room for the intervention of non-physical properties even in the processes of synaptic transmission. In what follows, however, I will just *assume* that the electrochemical properties physical sciences are familiar with are sufficient – and, more precisely, *immediately* sufficient – to account for neurotransmission. Still, I contend that something more seems to be needed if the ultimate purpose of the physicalist is to exclude the possibility of interactionist dualism.

Upon closer inspection, the fact that *sui generis* mental properties are not involved in the process of synaptic transmission is perfectly compatible with the claim that they do play a role in the production of our behaviour. As the story goes, dualists argue that *sui generis* mental properties emerge when a certain level of complexity is reached within the nervous system. Accordingly, they are not compelled to claim that *sui generis* mental properties exercise their powers within the processes of synaptic transmission, that serve as building blocks for the functioning of the whole. Such properties may well exercise their powers within the nervous system at a higher level of organisation.

That said, physicalists are arguably aware of the problem. When focusing on the mechanisms of synaptic transmission, they seem to resort to an implicit further premise, namely that the behaviour of the nervous system is *compositionally determined* by the behaviour of its working units, that are neurotransmission mechanisms.[27] In other terms, the behaviour of the nervous system would be *completely* determined by (i) the behaviour of its components and (ii) the way these components are spatiotemporally organised. Note that this is not the same as saying that the behaviour of the nervous system's organised components is *nomologically* sufficient for the behaviour of the whole. This would leave room for the non-overdetermining contribution of *sui generis* emergent mental properties, that are usually taken to be nomologically dependent on the physical configurations they emerge from. If an analysis in terms of nomological sufficiency were to be provided, then the physicalist should make explicit that the involved natural laws are physical ones. In this case, the transitivity of nomological sufficiency would be blocked. Physical laws do not

[27] I am grateful to David Papineau for pointing out this to me.

account for the emergence of *sui generis* mental properties, that is governed by *special* natural laws.[28]

Avoiding unnecessary complications, the point is that the physicalist takes the behaviour of the nervous system to be the sum of the behaviours of the neurons – and glial cells, and so forth – it is made up of, and nothing else. Clearly, once this premise is brought into the picture, we can easily infer CC. If we are justified in holding Synaptic-CC and we have reasons to think that all causal processes within the nervous system are nothing but sums of mechanisms of synaptic transmission, then it follows that we are justified in believing that all physical events within the nervous system have immediately sufficient physical causes.

Framing the discussion in terms of combination principles can be useful. I have conceded that the physicalist can rely on a fairly detailed understanding of the physical causes involved in neurotransmission processes. However, this understanding would be useless without some insights into the way these processes combine. In order to take the evidence for Synaptic-CC to be evidence for CC as well, the physicalist takes for granted an *additive* principle of composition concerning the internal causal organisation of the nervous system:

> (Additivity) powers in combination produce the sum of the manifestations they produce independently[29]

Some specifications are necessary. Consider the case of a watch. When compared to the nervous system, it turns out to be a fairly simple mechanism. Even in such a case, however, it would be trivial to point out that the whole does something that its components, as well as the mere sum of them, cannot do. A hand alone does not tell the time, nor does a heap of gears randomly combined. Clearly, the components of the watch have to be spatiotemporally organised in an appropriate fashion. This is arguably valid for the greatest majority of the systems one could consider, even for the least complex ones. In what follows, I will not further discuss this issue, and I will take the appropriate-organisation clause to be implicit in Additivity.

At this stage, unsurprisingly, it all comes down to the following question: when it comes to the nervous system, is there evidence – be it theoretical or empirical – for Additivity or some similar combination principles that allow for the inference from Synaptic-CC to CC? If not, then the argument from physiology fails to vindicate a version of the causal closure principle that is strong enough to rule out the possibility of causally efficacious emergent mental properties.

Before concluding by discussing a possible strategy the physicalist could resort to, let me briefly make a related point. If one thinks about it, Additivity is exactly what strongly emergent properties violate (Robb 2018). Let us consider the characterisation of emergent properties as properties that 'confer causal capacities on the object that go beyond the summation of capacities directly conferred by the object's microstructure' (O'Connor and Wong 2005, p. 665). Now, suppose that Additivity

[28] On this point, see Yates (2009).

[29] I borrow this formulation from Robb (2018). Note that Robb is not committed to the principle.

was somehow vindicated with respect to the nervous system. The physicalist would have *direct* reasons against strong emergentism about mental properties, and it is unclear whether the causal argument would still have a role in the dialectic between the physicalist and the interactionist dualist. In other words, there seems to be a sense in which evidence for Additivity would be crucial in the economy of the causal argument, since it would allow for the inference from Synaptic-CC to CC. However, this evidence would also make the causal argument somewhat redundant. Were Additivity vindicated, interactionist dualism would be already out of the picture.

That said, let us leave this issue aside and focus on a possible way to support Additivity. Vindicating the claim that the internal causal structure of the nervous system is additive is far from being an easy task. As far as I can see, the best way to do it requires two steps:

1. identifying the causal profiles of the individual working units of the nervous system and the physical laws that govern their interactions;
2. once we have *complete* knowledge about them, checking whether the occurrence of the physical effects that are usually supposed to have mental causes follows.

If the answer to (2) is positive, then Additivity – and therefore CC – is vindicated. Crucially, this is not the same as proving that *sui generis* mental properties have no causal power at all. As we have seen, the possibility of overdetermination is compatible with CC holding.

Unfortunately for the physicalist, even if the problem of the calculation's difficulties is left aside, this strategy is hardly viable. In particular, a couple of points are worth highlighting. Admittedly, it is difficult to deny that, if our behavioural responses follow from the summation of the (organised) powers of the physical micro-components of the nervous system, then *sui generis* mental properties could be either causally inefficacious or at best genuinely overdetermining. However, things get complicated if (2) is answered negatively. In principle, this could count as evidence for the emergentist view that *sui generis* mental properties play an ineliminable role in the production of our behaviour. Physicalists, however, would arguably resist such a conclusion. More likely, they would argue that we have simply failed to identify some physical micro-powers or laws. Note that, in doing so, the physicalist would not be stubbornly begging the question. Indeed, (2) requires that we have *complete* knowledge of the powers of the nervous system's working units and the relevant physical laws governing their interactions. However, it is far from clear that we can come to know, at some point, that our knowledge about them is complete.

So far so good, at least for the physicalist. If the answer to (2) is positive, Additivity is vindicated. If the answer is negative, Additivity is still a possibility, although not an empirically supported one. The real issue with the strategy in question is that, once again, it is extremely demanding in terms of knowledge of the microphysical processes within our nervous system. Our brain is an incredibly complex object, to say the least, made up of 86 billion neurons connected by a huge number of synapses. Clearly, it is not the case that our brain is *entirely* responsible for each of our behavioural responses. Still, the size and the intricacy of the nervous

activations that are supposed to be causally responsible for our behaviours make (1) an utterly unrealistic goal to achieve, at least given the current state of research and the tools we have available. True, the study of the physiology and the functioning of the brain has made great progress in the last decades. Still, we are far from knowing the precise causal role of all the individual micro-components of the nervous system, and we largely ignore how they work together when it comes to bringing about our behaviour. Maybe, in the future, neurophysiology will provide detailed, microscopical descriptions of the causal processes taking place within our nervous system. Currently, however, this is nothing more than an (extremely) optimistic expectation concerning the development of brain science. Unfortunately for the physicalist, this is not enough for vindicating Additivity.

Taking stock, the physicalist's move to insist on the mechanisms of synaptic transmission is promising, at least prima facie. As we have seen, there is a considerable amount of empirical evidence suggesting that neurotransmission mechanisms involve only physical causes that are immediately sufficient for their effects. The problem is that, to legitimately generalise this up to CC, the physicalist should be in the position to take for granted something along the lines of Additivity with respect to the nervous system. Vindicating the claim that the micro-powers of the nervous system combine in an additive fashion, however, is not an easy task. I have briefly discussed what seems to be the most straightforward way to do it. However, resorting to the two-step strategy I have outlined is clearly not an option, at least at the moment.

Acknowledgments I am grateful to David Papineau, Alfredo Tomasetta and Michele Di Francesco for the numerous discussions on the causal argument. They have also provided me with extremely helpful comments and feedback on various drafts of this work. An earlier version of this paper was presented at the tenth edition of the Research Workshop on Philosophy of Biology and Cognitive Science (PBCS X). I would like to thank the audience for their questions and suggestions. Finally, I am grateful to two anonymous reviewers for providing extensive and helpful comments.

References

Alter T (2021) A defense of the supervenience requirement on physicalism. Thought: A J Philos 10(4):264–274

Baumeister RF, Lau S, Maranges HM, Clark CJ (2018) On the necessity of consciousness for sophisticated human action. Front Psychol 9:1925

Baysan U (2020) Causal emergence and epiphenomenal emergence. Erkenntnis 85(4):891–904

Bennett K (2003) Why the exclusion problem seems intractable, and how, just maybe, to tract it. Noûs 37(3):471–497

Bernstein S (2016) Overdetermination underdetermined. Erkenntnis 81(1):17–40

Bruntrup G, Jaskolla L (eds) (2016) Panpsychism: contemporary perspectives. Oxford University Press, New York

Campbell K (1970) Body and mind. Macmillan, London; Toronto

Chalmers DJ (1996) The conscious mind: In search of a fundamental theory. Oxford University Press, Oxford

Chalmers DJ (2006) Strong and weak emergence. In: Clayton P, Davies P (eds) The re-emergence of emergence. Oxford University Press, New York, pp 244–255
Chalmers DJ (2010) The character of consciousness. Oxford University Press, New York
Crane T, Mellor DH (1990) There is no question of physicalism. Mind 99(394):185–206
Craver CF (2007) Explaining the brain: mechanisms and the mosaic Unity of neuroscience. Oxford University Press, New York
Dasgupta S (2014) The possibility of physicalism. J Philos 111(9/10):557–592
De Brigard F (2014) Self-stultification objection. J Conscious Stud 21(5–6):120–130
Dennett DC (1991) Consciousness explained. Little, Brown and Company, Boston
Di Francesco M, Tomasetta A (2015) The end of the world? Mental causation, explanation and metaphysics. Humana Mente 8(29):167–190
Dimitrijević DR (2020) Causal closure of the physical, mental causation, and physics. Eur J Philos Sci 10:1
Fodor JA (1989) Making mind matter more. Philos Top 17(1):59–79
Garcia R (2014) Closing in on causal closure. J Conscious Stud 21(1–2):96–109
Gibb S (2010) Closure principles and the laws of conservation of energy and momentum. Dialectica 64(3):363–384
Gibb S (2015) The causal closure principle. Philos Q 65(261):626–647
Goff P (2017) Consciousness and fundamental reality. Oxford University Press, New York
Gomes G (2002) The interpretation of Libet's results on the timing of conscious events: a commentary. Conscious Cogn 11(2):221–230
Jackson F (1982) Epiphenomenal Qualia. Philos Q 32(127):127–136
James W (1879) Are we automata? Mind 4:1–22
Kim J (1976) Events as property exemplifications. In: Brand M, Walton D (eds) Action theory. Springer, Dordrecht, pp 159–177
Kim J (1996) Philosophy of mind. Westview Press, Boulder
Kim J (1998) Mind in a physical world. An essay on the mind-body problem. MIT Press, Cambridge, MA
Kim J (1999) Making sense of emergence. Philos Stud 95:3–36
Kim J (2005) Physicalism, or something near enough. Princeton University Press, Princeton; Oxford
Koch C (2019) The feeling of life itself: why consciousness is widespread but can't be computed. MIT Press, Cambridge
Lavazza A (2016) Free will and neuroscience: from explaining freedom away to new ways of operationalizing and measuring it. Front Hum Neurosci 10:262
Lewis D (1983) New work for a theory of universals. Australas J Philos 61(4):343–377
Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. Behav Brain Sci 8(4):529–566
Lowe EJ (2000) Causal closure principles and emergentism. Philosophy 75(294):571–585
Lowe EJ (2003) Physical causal closure and the invisibility of mental causation. In: Walter S, Heckmann H (eds) Physicalism and mental causation: the metaphysics of mind and action. Imprint Academic, Exeter, pp 58–78
McLaughlin BP (1992) The rise and fall of British emergentism. In: Beckerman A, Flohr H, Kim J (eds) Emergence or reduction? Essays on the prospects of non-reductive physicalism. De Gruyter, Berlin, pp 49–93
Mele AR (2014) Free: why science hasn't disproved free will. Oxford University Press, New York
Melnyk A (2006) Realization and the formulation of physicalism. Philos Stud 131(1):127–155
Montero BG (2003) Varieties of causal closure. In: Walter S, Heckmann H (eds) Physicalism and mental causation: the metaphysics of mind and action. Imprint Academic, Exeter, pp 173–187
Montero BG (2013) Must physicalism imply the supervenience of the mental on the physical? J Philos 110(2):93–110
Montero BG, Brown C (2018) Making room for a this-worldly physicalism. Topoi 37(3):523–532

Montero BG, Papineau D (2005) A defence of the via negativa argument for physicalism. Analysis 65(3):233–237
Montero BG, Papineau D (2016) Naturalism and physicalism. In: Clark KJ (ed) The Blackwell companion to naturalism. Wiley, Malden and Oxford, pp 182–195
Noordhof P (2010) Emergent causation and property causation. In: Macdonald C, Macdonald GF (eds) Emergence in mind. Oxford University Press, New York, pp 69–99
O'Connor T (1994) Emergent properties. Am Philos Q 31(2):91–104
O'Connor T, Wong HY (2005) The metaphysics of emergence. Noûs 39(4):658–678
Owen M (2020) The causal efficacy of consciousness. Entropy 22(8):823
Papineau D (1998) Mind the gap. Philos Perspect 12:373–388
Papineau D (2001) The rise of physicalism. In: Gillet C, Lower B (eds) Physicalism and its discontents. Cambridge University Press, Cambridge, pp 3–36
Papineau D (2002) Thinking about consciousness. Oxford University Press, New York
Papineau D (2009) The causal closure of the physical and naturalism. In: Beckermann A, McLaughlin BP, Walter S (eds) The Oxford handbook of philosophy of mind. Oxford University Press, New York, pp 53–65
Papineau D (2020) The problem of consciousness. In: Kriegel U (ed) The Oxford handbook of the philosophy of consciousness. Oxford University Press, New York, pp 14–38
Pereboom D (2011) Consciousness and the prospects of physicalism. Oxford University Press, Oxford
Pockett S, Purdy SC (2011) Are voluntary movements initiated preconsciously? The relationship between readiness potentials, urges and decisions. In: Sinnott-Armstrong W, Nadel L (eds) Conscious will and responsibility. Oxford University Press, New York, pp 34–46
Popper KL, Eccles JC (1977) The self and its brain. Springer, Berlin
Robb D (2018) Could mental causation be invisible? In: Carruth A, Gibb S, Heil J (eds) Ontology, modality and mind: themes from the metaphysics of E. J. Lowe. Oxford University Press, New York, pp 165–176
Robinson WS (2019) Epiphenomenal mind: an integrated outlook on sensations, beliefs, and pleasure. Routledge, New York; London
Shapiro L (ed) (2007) The correspondence between Princess Elisabeth of Bohemia and Descartes. The University of Chicago Press, Chicago and London
Sider T (2003) What's so bad about overdetermination. Philos Phenomenol Res 67(3):719–726
Tomasetta A (2015) Physicalist naturalism in the philosophy of mind. Discipline Filosofiche 25(1): 89–111
Wilson J (2015) Metaphysical emergence: weak and strong. In: Bigaj T, Wüthrich C (eds) Metaphysics in contemporary physics. Brill, Leiden and Boston, pp 345–402
Wilson J (2021) Metaphysical emergence. Oxford University Press, New York
Yates D (2009) Emergence, downwards causation and the completeness of physics. Philos Q 59(234):110–131
Zargar Z, Azadegan E, Nabavi L (2020) Should methodological naturalists commit to metaphysical naturalism? J Gen Philos Sci 51:185–193

Color and Competence: A New View of Color Perception

Tiina Rosenqvist

Abstract I have two main goals in this paper. My first goal is to sketch a new view of color perception. The core of the view can be expressed in the following two theses: (i) the overarching *function* of color vision is to enable and enhance the manifestation of relevant (species-specific) competences and (ii) color experiences are *correct* when they result from processing that directly and non-accidentally subserves the manifestation of such competences. My second goal is to show that the view can accommodate and account for a wide variety of color perceptual phenomena, including many problem cases. Importantly, the framework allows us to differentiate between two kinds of good cases of color perception: *ideal* cases where the demands of the relevant competences converge and *non-ideal* cases where the demands of the relevant competences diverge and clash.

Keywords Color perception · Competence · Correctness · Function · Color illusion

1 Introduction: Why Non-Ideal Cases Matter from the Start

Philosophical views of color are often inspired by the most straightforward cases of correct color perception. Call these "ideal" cases. An example of an ideal case would be the light bouncing off the surface of a ripe tomato causing a human trichromat with a normally functioning visual system to experience redness where the tomato is in the visual field. There is nothing pathological about this case, nothing abnormal about the causal chain, and nothing too surprising about the percept itself.

We want a compelling story of what occurs when we perceive ripe tomatoes as red. At the same time, however, if we let ideal cases drive our inquiry, we run the risk of oversimplifying both our explanandum and our explanans. To see why this is, consider the following scenario. Say Adrian is a budding cartographer in search of an

T. Rosenqvist (✉)
Department of Philosophy, University of Pennsylvania, Philadelphia, PA, USA
e-mail: trosenq@sas.upenn.edu

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_5

account of what makes a map good. She lives in a small desert, surrounded by just a handful of objects: some large rocks, a couple of houses, and a few cactus plants. When Adrian constructs her theory of good map-making, she does not consider any other kinds of environments, and concludes that a good map (i) depicts all objects bigger than a grain of sand that a normal human perceiver situated in the mapped location could effortlessly perceive in good light, and (ii) accurately represents the spatial relations that obtain between those objects. It would be easy enough to create a map of this sort of Adrian's desert; one could even draw a grid in the sand and a corresponding grid on paper to find where objects should go (in this sense, Adrian's desert is an ideal terrain for map-making purposes). It is clear, however, that Adrian's standards are much too high for maps in general. Most perfectly good maps, no matter how detailed and carefully crafted, do not depict all objects bigger than a grain of sand.

The cartography example helps illustrate the dangers involved in heavy reliance on ideal cases when philosophizing about color perception. A good cartographer often must compromise and make choices (say, by representing a tangle of trees with a single tree icon) to maximize the "goodness" of her maps. It is possible that color visual systems are like cartographers in this respect: when the perceptual situation is not straightforward or ideal, the color visual system too might have to compromise. A compromise need not mean that the system is not functioning well, or that the resulting color percept is somehow incorrect. It might just mean that we need a more sophisticated account of what makes color visual systems or specific instances of color perception successful.

I have two main goals in this paper. First, I wish to sketch a new view of color perception, one that is not inspired by a narrow focus on ideal cases and one that considers the many ways in which color vision can be useful to perceiving organisms. Second, I wish to show how the view can account for a wide variety of color perceptual phenomena, including many problem cases. I proceed as follows: in Sect. 2 I introduce three desiderata for philosophical theories of color perception. In Sect. 3 I offer a first-pass characterization of my view: I suggest that color perception is *competence-embedded* in the sense that its overarching function is to enable or enhance the manifestation of relevant competences, and that color experiences are correct when they result from competence-enhancing processing. In Sect. 4 I show how this framework suggests a plausible explanation for some "textbook color illusions" by allowing us to distinguish between two kinds of good cases of color perception: ideal and non-ideal. In Sect. 5 I consider some pertinent objections and discuss ways in which we might refine and develop the view in response. In Sect. 6 I conclude that the view not only comes with a lot of explanatory purchase, but also allows us to uphold the powerful intuition that color visual systems are generally well-functioning systems that nevertheless sometimes fail.

2 Three Desiderata for Philosophical Accounts of Color Perception

Before I characterize the view, it is helpful to consider some general features that we look for in a philosophical account of color perception. In this section I introduce three basic desiderata that are based on robust intuitions shared by many philosophers working on color. Although not wholly rigid and immutable, the desiderata mark a starting point for philosophizing about color perception and the sacrifice of any particular desideratum demands careful justification.

2.1 A View Should Not Attribute Widespread Failure to (Normally Functioning) Color Visual Systems

The first desideratum is based on a powerful intuition that color visual systems do their job fairly well. Humans tend to see color (or see *in* color) in ways that are consistent, coherent, and action-guiding. In addition, many other animals rely on color vision when they navigate, learn and memorize, and when they interpret or respond to their surroundings. It therefore seems intuitive to think that color visual systems are generally well-functioning systems.

A theory which entails widespread color misperception is in tension with this intuition.[1] An extreme example is provided by certain eliminativist philosophies of color (e.g., Hardin 1993; Maund 2006) which maintain that nothing in the external world is actually colored and that all color experiences are non-veridical. Implicit here is the idea that we experience color, think about color, and talk about color as if it were a stable (perhaps even mind-independent) property of physical objects, and that this is the kind of property that color would have to be for it to be *real*. Because we have good reasons to think that color is *not* this kind of a property, argues the eliminativist, we should conclude that all color experiences are illusory.[2]

Eliminativists face difficulties when it comes to explaining how color vision can be action-guiding and useful. If colors are illusory, then why do we experience them in the first place? How is a visual system that produces illusory experiences fitness-enhancing, and if it is not, then why did such systems evolve in multiple phylogenetic lineages? Some eliminativists have argued that what is required for usefulness is just reliability, and not correctness per se (Mendelovici 2016), but this strategy makes it difficult to account for the difference between useful and "useless" reliably occurring color experiences. Seeing an apple as red against green foliage is arguably

[1] For example, Byrne and Hilbert admit that the entailment "that many of us misperceive unique greens" is an "unwelcome result" (1997, p. 274). See also Gert (2006).

[2] This kind of eliminativism accepts an objectivist treatment of color, but eliminativism about the colors of external objects is also consistent with the view that colors are instantiated by mental objects instead.

useful, whereas seeing the spinning Benham disc – an object which normally appears black-and-white when stationary – as colored does not seem to be. If both cases are instances of *reliably* occurring misperception, then what explains the difference? As long as the eliminativist has no convincing story to tell here, we have good reason to resist the view.

Most philosophers working on color today are color realists. But the first desideratum also puts pressure on many realist theories, especially the kind that attribute stable, fine-grained (determinate) colors to objects and propose that veridical color perception consists in accurate detection and representation of those colors (e.g., Hilbert 1987; Tye 2000; Byrne and Hilbert 2003). The reason for this is simple: variation in color perception is systematic and extensive, and if we are to attribute *any* fine-grained colors to objects, it is difficult to escape the conclusion that most of our perceptual experiences deviate from the good case in which colors are correctly perceived.

2.2 A View Should Allow for Instances of Color Visual System Failure

The second desideratum is based on an equally powerful intuition that color visual systems do not *always* perform the way they are supposed to perform; that we *sometimes* see colors (or see *in* color) in ways that are confusing, inconsistent, or incorrect. We might hallucinate color or our color experiences could be distorted in some way. To accommodate this intuition, a philosophical view of color perception should allow for instances of incorrect color perception or at least explain our intuitions and ordinary discourse in some satisfactory way.[3]

If we think of color visual systems as having specific functions, and of color experiences as being correct when those functions are fulfilled, then incorrect color perception can be understood as color visual system failure. A theory of color perception should therefore allow for instances of color visual system failure where the system falls short of doing what it is *supposed* to be doing. It is fairly common to think of biological systems as having normative functions and of being susceptible of malfunction. A cardiovascular system is considered impaired when it fails to maintain adequate blood flow to different parts of the body and legs are considered to malfunction when they fail to support locomotion. Although there are divergent

[3]This idea is commonly accepted by philosophers working on color and color perception. Boghossian and Velleman write that "seeing something as red is the sort of thing that can be illusory or veridical" (1989, p. 82). Chirimuuta suggests that it is strongly intuitive to think that color misperception occurs (2015, p. 179). Cohen maintains that the "idea that there are errors of color perception is so fundamental to our (naïve and scientific) thinking about the visual system that it would be very difficult to accept a theory of color that failed to sustain it" (2007, p. 349).

philosophical analyses of the nature of normative functions,[4] for my purposes here it suffices to say that the notion of normative function is deeply-rooted in both ordinary linguistic and scientific practice, and is commonly applied to color vision as well. That is, we quite naturally think of color visual systems as goal-directed systems outputting experiences that are evaluable relative to that goal.

Unsurprisingly, there are competing conceptions of the function of color vision (e.g., Hatfield 1992; Hilbert 1992; Matthen 2005; Chirimuuta 2015), and different conceptions entail different degrees of color visual system failure. Most views uphold the basic intuition that some of our color experiences are incorrect, although there is a worry that some radically relativist views (McLaughlin 2003; Cohen 2009) fail to satisfy this desideratum. If colors are made relative to particular perceivers and particular perceptual circumstances, it is difficult to see how color perception could go wrong, except perhaps in instances of straightforward hallucination where nothing at all is really seen.

2.3 *A View Should Allow us to Evaluate and Explain Specific Color Experiences*

The third desideratum asks that a philosophical view of color perception provide the means to evaluate the correctness of color experiences and that it suggests robust philosophical explanations for various kinds of color phenomena. The first part of this desideratum is straightforward enough. Arguably, any philosophical account of color perception worth its salt has some way of adjudicating whether a particular chromatic experience is correct or incorrect.[5] If the verdict is placed behind an epistemic veil of ignorance, it is unclear what work the notion of correctness can do.

If a view cannot unambiguously adjudicate *when* color experiences go wrong, it might also be less equipped to explain *why* they go wrong. This connects to the second part of the desideratum. Consider intrapersonal variation due to changes in chromatic context, for example. Suppose we have a theory which says that each uniform surface instantiates just one stable, fine-grained color. Suppose also that our target surface (e.g., a paint sample card) gives rise to markedly different color

[4]The basic dividing line is between "Wrightian" (Wright 1973, 1976) etiological analyses of function that appeal to natural selection (e.g., Neander 2017) and Cummins-style (Cummins 1975) causal-role functional analyses which find normativity in the practices of the scientific community (e.g., Hardcastle 2002).

[5]For example, one of the main objections to traditional dispositionalism – the view that colors are dispositions to cause certain kind of chromatic experiences in normal perceivers in standard conditions (e.g., Levin 2000) – is that it cannot specify "normal" perceivers and "standard" conditions in any satisfactory, non-arbitrary manner (Hardin 1993). But this criticism does not just apply to dispositionalism, but to *all* theories that posit stable, determinate object colors (e.g., Allen 2016; Campbell 1993; Byrne and Hilbert 2003). For criticism of the idea of unknowable color facts, see, e.g., Cohen (2003).

experiences in some perfectly ordinary chromatic contexts (e.g., when held against walls of different colors). Not only is it very difficult to determine which one of these chromatic experiences is the correct one, but the variation itself appears rather mysterious. It seems reasonable to ask why our perception should deviate from the good case in such a systematic (and potentially useful) manner, and if a view has nothing to say in response, that seems like a genuine weakness.

Whereas the first two desiderata concern the question of what philosophical views of color perception should entail about color perception on a more *general* level, the third desideratum concerns the application of views to *particular* instances of color perception. It is likely to be the most controversial desideratum out of the three, because it is not obvious that most philosophers think that their views should have illuminating things to say about specific color perceptual phenomena.[6] At the same time, however, it seems relatively uncontroversial that a philosophical theory that helps us make better sense of our target phenomenon has at least some advantage over theories that leave us in the dark, all other things being equal.

3 The Competence-Embeddedness of Color Vision

With these desiderata in mind, I will now sketch my own account of color perception. I side with Akins (2001) in thinking that the primary function of color vision is to help organisms see *better* and see *more,* and I side with Chirimuuta (2015) in thinking that the correctness standards for color experiences are directly tied to this sort of usefulness. Chirimuuta calls her approach "perceptual pragmatism."[7] She

[6] Some philosophers appear to think that insofar as their view entails the illusoriness of a (type of) perceptual experience, then no other explanation for that (type of) experience is needed. For example, Tye suggests that simultaneous contrast effects in color perception are "illusions or normal misperceptions" that can be explicated in terms of "the workings of the visual system" (2000, pp. 154–156), but does not himself attempt to explicate the issue any further.

[7] I consider myself to be directly building on the work of Akins and Chirimuuta here, but my view also shares similarities with certain other naturalistic, action-oriented accounts of color perception that link the correctness of color experiences to species-specific functions of color vision (e.g., Thompson 1995; Hatfield 2003; Matthen 2005). For example, I am generally sympathetic to Matthen's thesis that sensory systems are "automatic sorting machines" that sort environmental objects "into classes according to how they should be treated for the purposes of physical manipulation and investigation" (2005, p. 8). That said, I think that Matthen's epistemology of color perception *overemphasizes* the role color vision plays in the building of a "stable record" of the properties of environmental objects. Matthen proposes, for example, that a color experience is incorrect if it disposes a perceiver to make mistaken inferences about the ripeness of fruit, e.g., when the perceiver "misclassifies a particular fruit as pale green, although in fact it meets the physical specification of the sensory category, *yellow*" (ibid., p. 208). I do not deny that we often use color-looks to make such inferences, but I maintain that an experience of a normally yellow-appearing fruit as pale green need not be incorrect, and that there can even be situations where experiencing the fruit this way constitutes the *best case scenario* for the perceiving organism. On the other hand, my approach has very little in common with certain other "pragmatist" views. One example is

rejects the notion that the correctness of a perceptual state is determined by its correspondence to states of affairs in the world, and she presents her alternative, utility-based epistemology as a *naturalized epistemology derived from the theoretical commitments of perceptual scientists* (2015, pp. 101–110). I am sympathetic to both her general approach and her claim that we should evaluate color experiences in terms of usefulness (and not correspondence), but I worry that the notion of usefulness alone is too vague to do the kind of normative and explanatory work that we would ideally want our philosophical accounts of color perception to do. For this reason, I propose an analysis of usefulness in terms of *competence-embeddedness*. I start by explicating what I mean by "competence" (and "capacity"). I then explain what I consider to be the relevant kind of competences that directly embed color vision and allow us distinguish between correct and incorrect color experiences. Finally, I explain why I think that we should understand color perception as competence-embedded processing rather than as a full-fledged perceptual competence in its own right.

3.1 Competences and Capacities

I find it helpful to think of perceptual activity as consisting in the exercise of perceptual *competences*. First, the notion of competence captures the apparent skillfulness of perception, i.e., the idea that perceptual systems and processes are often reliable and well-functioning, and enable meaningful interactions with environments. That said, those readers who are suspicious of the idea of perception as reliable might prefer to think of perceptual activity as consisting in the exercise of perceptual *capacities* instead, since the notion of capacity is neutral on the question of reliability. Second, the notion of competence (and the notion of capacity) comes with conceptual resources to account for the differences in the aims of different kinds of perceptual systems and processes. In the case of vision, it allows us to distinguish between visual competences proper and the kind of visual processing that is merely competence-embedded. I take color vision to be an example of the latter.

There are different philosophical accounts on what competences (Greco 2007; Sosa 2015; Miracchi 2015) or capacities (Schellenberg 2018; Hornett 2021) are, but most of what I say in this section is going to be neutral between these accounts. For the sake of exposition, however, I am going to briefly describe Sosa's view. To start, it appears that animals can have various kinds of aims. Some of these aims are purely epistemic or cognitive, some are behavioral or agential, and some are perceptual or perceptual-cognitive. Sosa maintains that we can assess all aim-involving

Gert's neo-pragmatist account which takes color language as its starting point and aims to explain "how and why we talk the way we do" (2018, p. 225). Whereas *perceptual* pragmatism often starts with the question of the function of color vision and looks to vision science for help, Gert's *linguistic* pragmatism starts with the question of the function of color terms in ordinary discourse.

performances for accuracy, adroitness, and aptness (2007, pp. 22–23). When a performance is accurate, it achieves its aim. When a performance is adroit, it is produced by the exercise of the relevant competence. When a performance is apt, it is accurate because it is adroit, i.e., the success of the performance manifests the competence of the performer. Sosa's competences are types of dispositions, i.e., dispositions to succeed with aims, to perform well (2015, pp. 26–27). For example, if Bailey is a competent cyclist, she possesses a cycling skill that allows her to succeed in (safely) riding a bicycle in a certain range of internal and external conditions. If her performance manifests her cycling competence, its success owes to her exercise of the competence. Similarly in the case of visual perception, Bailey's visual performance manifests visual competence(s) when the visual images she hosts "aptly correspond" to the object with which she perceptually interacts (see ibid., p. 21).

Regardless of how we wish to understand competences/capacities, it is often intuitive enough to judge whether a competence/capacity is possessed. For example, it seems reasonable to think that many humans and bats are competent perceivers of distances, at least at the scale required for successful action. For my purposes here, it is important to emphasize that there are different ways to exercise a distance perception competence. Whereas some bats use biological sonar and auditory processing, humans tend to rely on visual processing. But if we go more fine-grained, we can see that even our own *visual* distance perception competence can be exercised in different ways. In low light, when only our rod photoreceptors are active, the exercise of our competence relies on luminance vision alone. But at higher levels of light most of us have an additional tool at our disposal – *color vision*.[8]

3.2 Relevant Competences

To recap, I think that the aim of color vision is to help the perceiving organism see better in general, and that this aim determines the correctness standards for color experiences. I am now ready to connect this idea to the notion of

[8] Humans have two kinds of retinal photoreceptors: rods and cones. Both absorb light as a function of wavelength and intensity, but whereas the rods all have the same wavelength sensitivity, cones come in different types (normal human perceivers have three classes of cones with absorption maxima in the short-, medium-, and long-wavelength regions of the visible spectrum). Color vision requires the comparing of the activity of at least two classes of receptors with different wavelength sensitivities. Because there is only a single type of photoreceptor active in very low light, rod-mediated vision is achromatic. At higher levels of light, cones become active and their outputs "are combined at the post-receptor level in two different ways: one additive, giving rise to luminance signals with no information regarding the wavelength composition of light, and one subtractive, which preserves the latter and can thus be used for determining the color of objects" (Moutoussis 2015, p. 5).

competence-embeddedness. When I say that color vision is competence-embedded (CE), I am making two interconnected claims:

> (CE-F) The overarching *function* of color vision is to enable and enhance the manifestation of relevant (species-specific) competences
>
> (CE-C) Color experiences are *correct* when they result from processing that directly and non-accidentally subserves the manifestation of relevant (species-specific) competences

CE-F is a claim about the *proper functioning* of the color visual system.[9] It specifies the normative function of color vision at its most fundamental and universal: color vision is *for* exercising competences, its job is to help an organism succeed with certain aims. The specific competences that color vision subserves might vary from species to species and it is an interesting empirical question what the species-specific relevant competences are and how they are organized. The claim I am making here is a more general one, however.

CE-C is an epistemological claim about the success conditions of color experiences. Although I think that the claim can be made for color *perceptions* more generally, my focus in this paper is on phenomenal experience. Ideally, a philosophical view of color perception tells us clearly and unambiguously when our color experiences are correct and when they are incorrect. My claim is that color experiences are correct when they result from processing that helps an organism manifest some relevant competence(s). When this is not the case, the experience can be said to be incorrect.

It seems natural to think that the competences relevant to the epistemology of color perception are competences that color vision *directly* and *non-accidentally* subserves. This condition helps keep various idiosyncratic interests and goals from making otherwise useless color experiences correct. For example, if I see a red afterimage which reminds me of my intention to buy apples, the experience of redness helps me manifest a competence in a purely accidental manner. My color visual system is not "aiming" to enhance or enable the manifestation of my competence to remember the items of my grocery list.

The competences that color vision directly and non-accidentally subserves are competences that color vision "aims" to subserve. At least some such competences are likely to have been fitness-enhancing in the animal's evolutionary environment. Color vision (or a particular type of color vision, e.g., a specific type of trichromacy) could then have been selected because it contributed to the manifestation of these fitness-enhancing competences. The analysis of the species-specific aims of color

[9] I have chosen to discuss the "color visual system" as if it were a unified entity and to lump together different kinds of processing that subserve different competences. Some readers might consider this an oversimplification and prefer instead to differentiate between two (or more) separate systems *within* color vision, e.g., one system that computes chromatic contrast and another that computes (more or less constant) surface color representations (see, e.g., Akins and Hahn 2014; Moutoussis 2015). In the end, nothing I say here requires the adoption of the first conception, and those in favor of the multi-system conception can read me as suggesting that color vision consists in the operation of different systems that subserve different perceptual competences.

vision can therefore benefit from etiological analysis: we can ask how a specific kind of color visual system might have contributed to its own persistence by contributing to the persistence of the visual system as a whole and, ultimately, to the persistence of the organism as a whole. For example, we can ask why trichromacy re-evolved in primates. If it evolved to aid scene segmentation and object recognition, for example, then we have good reason to think that scene segmentation and object recognition should be on our list of the relevant competences that embed color vision in trichromatic primates.[10]

That said, it is difficult to prove that primate trichromacy evolved to serve specific competences, and only those competences.[11] And even if this could be done, it might be that trichromacy has since acquired other functions, and that these functions have contributed to its maintenance and persistence. In other words, a particular type of color vision might have come to enhance competences that it did not originally evolve to enhance.[12] To get a more comprehensive understanding of the functional profile of color vision in a given species or population, the perceptual pragmatist looks to uncover the *current* aims of color visual systems by paying close attention to what scientists are saying about the functional organization of the visual brain and about the role that color vision plays in the perceptual economy of the organism. Visual ecology can provide clues as to how animals use color vision to achieve important behavioral goals. Psychophysics can shed light on the rules and principles that govern color vision, and these can then be used to draw inferences about the aims of the color visual system in question. Neuroscience can reveal the involvement of the color visual system in different kinds of perceptual and cognitive tasks, and these findings can then be used to ground function attribution. In humans, empirical research has already helped identify a number of candidate competences; we have good reason to think that color vision functions to enhance and enable depth perception, distance perception, shape and form perception, shadow perception, figure-ground segregation and scene segmentation, object identification and re-identification, property identification, and memorization.[13]

[10] On the origins and aims of primate color vision, see e.g., Jacobs (1981), Mollon (1989), and Dominy and Lucas (2001).

[11] We should generally exercise caution when proposing adaptive explanations for the number of cone types or the spectral tuning of those cones in a given species. As Chittka and Briscoe (2001) remind us, some sensory traits are better explained by phylogenetic constraint, evolutionary inertia, or random processes.

[12] Some etiological theorists require that proper functions reflect *recent* natural selection (e.g., Godfrey-Smith 1994), while others appeal to the "continuing usefulness" of traits selected for specific purposes (e.g., Schwartz 2002).

[13] See, e.g., Kingdom (2008), Shevell and Kingdom (2008), Troscianko et al. (1991), Smithson (2015), Tanaka et al. (2001), Paramei and van Leeuwen (2016), Gegenfurtner and Rieger (2000), and Moutoussis (2015). For a helpful overview, see Chirimuuta (2015, Chap. 4). Vision scientists themselves often engage in function attribution and seem sensitive to the distinction between the kind of perceptual phenomena that plausibly reflect the proper functions of color vision and the kind of perceptual phenomena that are mere by-products of the mechanisms of color vision.

The competences listed above are all "ecological" competences that enable crucial animal-environment interactions by allowing animals to detect, locate, track, identify, categorize, and remember ecologically important objects, object properties, and relations. To be sure, color vision can play a part in the manifestation of other kinds of competences as well. For example, paying close attention to the chromatic appearance of a painting could have helped ascertain its value in 15th century Italy (see Baxandall 1988). But note that color vision could only help manifest such painting appraisal competence by first enhancing or enabling the manifestation of a competence to identify the use of certain expensive pigments, such as lapis lazuli-derived ultramarine. In other words, color vision could only help manifest the non-ecological competence indirectly, via the manifestation of a more basic ecological competence: property identification.[14]

What ultimately determines the correctness of a particular color experience, then, is the relevant species-specific ecological competences. As the red afterimage example shows, sometimes color-involving experiences can be useful without being correct. This statement can now be understood as referring to cases where color vision is useful without helping the perceiver manifest any relevant ecological competences.

3.3 Case Study: The Color Visual System and Figure-Ground Segregation

I will now consider an example of a relevant ecological competence in humans: figure-ground segregation. This is an essential step in visual processing in which individual objects come to be perceived as figures bounded by closed contours. Separating objects from backgrounds requires that a border shared by two visual regions be assigned to one of those regions, producing a percept of a shaped object located closer to the perceiver.[15]

Figure-ground segregation is rarely a conscious aim. It tends to be an automatic process to which we rarely pay attention unless it becomes challenging for some reason, e.g., due to poor visibility. This is not to say that the problem is trivial. A number of features of natural scenes make it a challenging task: objects are often partially occluded, and scenes tend to be cluttered and noisy, e.g., incident light can create false boundaries and certain types of camouflage can make objects blend with their backgrounds. Thus, when assigning boundaries to objects, visual systems often

[14] Of course, the ability to recognize ground-up lapis lazuli from the way it chromatically appears likely had nothing to do with the evolution of color vision in humans and other primates. But this ability is a special case of a more general, ecologically-relevant ability to recognize properties (being ripe, being angry, etc.) from the way things chromatically appear (red, etc.).

[15] These are objects at the relevant, species-specific level of description. For humans, this means things like apples, tables, and mosquitoes.

must rely on a number of factors ranging from low-level image-based cues (convexity, symmetry, small area, closure, top-bottom polarity, etc.) to higher-level factors such as past experience.[16]

Phenomenologically speaking, meaningful objects are defined both by their boundaries and by their surfaces. When I perceive a cat napping on an armchair, I perceive the boundary of the cat as a discontinuity in space, and the surface of the cat as a continuous region to which the boundary belongs. Some have suggested that boundary detection alone is sufficient for figure-ground segregation (e.g., Biederman 1987), but there exists compelling evidence that figure-ground segregation is influenced by, and sometimes requires, surface representation (see, e.g., Smithson 2015; Yamagishi and Melara 2001). An interesting question for my purposes is *if* and *how* color vision contributes to figure-ground segregation. Does it contribute to surface representation, to boundary representation, to neither, or to both?

There is plenty of evidence of the involvement of the color visual system in surface representation (see, e.g., Dresp-Langley and Reeves 2014; Palmer and Brooks 2008). De Valois and Switkes find asymmetric interaction between (isoluminant) chromatic stimuli and (isochromatic) luminance stimuli; whereas chromatic gratings appear to profoundly mask luminance gratings, the converse is not true. The authors offer an ecological explanation: because there normally exists a great deal of luminance noise in our environments, "it would be an organism's benefit to segregate objects from the backgrounds on the basis of color differences rather than luminance differences" (1983, p. 17). That said, the received view in perceptual psychology and neuroscience has been that contour perception is achieved by the luminance system. For example, Rogers-Ramachandran and Ramachandran (1998) propose that an "essentially color-blind," fast-acting system extracts visual contours, whereas a slower system computes surface colors. But this suggestion now appears oversimplified, and a growing body of evidence points to the important role that color vision plays in the perception of object boundaries.[17] Neuroscientists have identified "color-sensitive," orientation-tuned neurons in the primate visual cortex that respond to pure chromatic gratings (see Shapley et al. 2014 for a review)[18] and psychophysicists have confirmed the contributions of color vision to primate contour perception (see, e.g., Moutoussis 2015; Hansen and Gegenfurtner 2017).

Overall, we have very good reason to think that our color visual system helps with figure-ground segregation by contributing to both contour representation and surface

[16] For a review, see Peterson (2015).

[17] When it comes to boundary computations, the cells in the luminance system respond to luminance edges even in the absence of chromatic contrast, and the cells in the color visual system respond to (certain kinds of) chromatic edges even in the absence of luminance contrast. Therefore, the luminance system and the color visual system need not be thought of as being completely independent and modular. There is likely to be interaction between the two and the systems might even share some neural resources in the visual cortex (see, e.g., Shapley et al. 2014, p. 577; Moutoussis 2015, p. 6).

[18] Garg et al. (2019) suggest that even the majority of color-*preferring* neurons in the primary visual cortex might be strongly tuned for orientation.

representation. That both the color visual system and the luminance system should enhance and enable the manifestation of this very important perceptual competence is hardly surprising. As Akins and Hahn remind us, "when a sensory system guides the real-time behaviour of an organism, any information that makes its visual computations faster, cheaper, or more reliable is ripe for selection" (2014, p. 139). In other words, the improved efficiency and reliability of such computations alone seems to explain why color visual systems exist in many different types of animals.

That said, we still need to address one important question. So far, I have suggested that the primary goal of color visual systems is to enable and enhance the manifestation of relevant competences. But why should we think that color perception is merely competence-embedded, and not a perceptual competence in its own right? In other words, why is it not the *primary* aim of color vision to see colors, and its competence-enhancement role merely secondary?

3.4 Why Color Perception Is Not a Competence

To argue that color perception is competence-embedded in my sense is to argue that color vision aims to enable and enhance the manifestation of relevant competences, and that *this enhancement function fully determines its success conditions.*

If color perception were itself a competence/capacity, then what would be its function? A natural answer is that the function would be to perceive (to detect, represent, and/or successfully engage with) the chromatic properties of our environments. This, in turn, suggests that the success conditions of color perception would have to do with *correspondence:* a color experience would be successful if it matched the chromatic property instantiated by the object (and if the success resulted from an exercise of a competence/capacity), and unsuccessful if it did not.

This fits in well with how Schellenberg defines the function of a perceptual *capacity*. She suggests that "the function of a perceptual capacity C_α is to discriminate and single out mind-independent particulars α_1, α_2, α_3, ... α_n, that is, particulars of a specific type" (2018, p. 34). In the case of color perception, this would mean that there had to exist some stable objective chromatic properties for the color visual system to discriminate and single out. For example, the function of the capacity to perceive red would be to discriminate and single out red particulars in the world. This function would be fulfilled in veridical perception, whereas the same capacity would be unsuccessfully employed in the case of non-veridical perceptual experience (see ibid., p. 43).[19]

[19]The idea is that when we hallucinate a red particular, a red particular does not *really* exist to be discriminated and singled out. When our perception of a given particular as red is illusory, the particular exists but does not instantiate the property of redness. In both cases, then, there is a lack of correspondence between the perceiver's perceptual state and the state of the world.

Miracchi (2017) offers an alternative account of perceptual competences. She does not require that the properties we engage with be objective and mind-independent, but the idea of stability plays a role in her approach as well. This is evident in her requirement that "appropriate regularities" involving the agent and the perceived object obtain, and that "perceptual competences reliably issue in cases of perceiving things as they are" (2017, pp. 645, 650). In the case of color perception, a natural way to understand such regularities would be to link a particular (type of) object surface to a particular (type of) chromatic experience in a given perceiver. For example, the manifestation of a competence to perceive things to be red would require that the object in question were, in fact, red (i.e., reliably linked to red experiences in that perceiver). Perceptual illusion, on the other hand, would occur when an object reliably linked to, *say*, blue experiences were experienced as red instead (see ibid., p. 663).

Although very different, Schellenberg's and Miracchi's accounts both emphasize some sort of a stable link between perceptual experiences and the objects being perceived, and conceptualizing color perception as a competence/capacity seems to naturally go with an emphasis on the stability and constancy of color experience.[20] The problem with this is that color visual systems also appear to have "aims" that do not require such stability; the contribution of the color visual system to contour perception is a case in point. Notice also that variation in color perception is systematic and rife, even on an intrapersonal level. Our phenomenal experiences of the colors of surfaces (etc.) are influenced by such factors as lighting conditions, chromatic contexts, viewing angles, and viewing distances. Although this kind of variation could be conceptualized as color visual system failure, it is often *useful* to the perceiver. For example, if a ripe, normally red-appearing apple looks even redder when surrounded by green foliage, this increases the conspicuousness of the apple in that specific context. And if our perceptual engagement with a target smooths out some of its (physical) surface variation through the process of color assimilation, this plausibly enhances the perceived figureness of the object. Because these kinds of effects are incredibly common in everyday perception, and because plausible ecological explanations for the effects exist, it seems perfectly reasonable to assume that the effects might reflect the aims of color visual systems and feature in perfectly good cases of color perception.

Recall that empirically-oriented philosophers such as Akins contend that color vision is not for seeing colors or for seeing things as colored, but for seeing better in general. If you add to this Chirimuuta's view that the success standards of color experiences are entirely utility-based, then the resulting views have no use for the kind of regularities between objects and perceptual experiences that competence/capacity accounts of color perception rely on to get off the ground. The best way to

[20] One could attempt to formulate a radically relativist competence account of color perception by proposing that the appropriate regularities obtain between particular perceptual agents, object surfaces, and *particular perceptual circumstances*. Although this would eradicate the need to posit stable object color, it would also dilute the notion of perceptual competence.

make a competence framework compatible with this sort of pragmatism is to hold that color perception is competence-embedded, rather than a competence itself. But this is precisely why the notion of competence is so useful. It helps us make sense of the idea that not all the properties presented to us in visual experience are on a par; some are *what* we (can) competently perceive, others are *how* we (can) competently perceive.

4 "Textbook Color Illusions" as Test Cases

"Textbook color illusions" are images created to illustrate systematic color perceptual phenomena. The relevant phenomena are often subtle enough to escape our attention in everyday contexts, and the images are designed to increase the magnitude of the effects to make them more noticeable. In this section I will use two such "illusions" to test the explanatory potential of CE. Both are examples of color induction where a neighboring region induces a shift in the perceived chromaticity of a target region. I suggest that the perception of these and many other textbook illusions constitute paradigmatic "non-ideal" cases of color perception where the demands of the relevant competences clash.

4.1 Clashing Competences

Thus far, I have suggested that the function of color vision is to enable and enhance the manifestation of relevant competences. Now I want to employ the notion of competence-embeddedness to distinguish between two kinds of successful or "good" cases of color perception: *ideal* and *non-ideal.*

I start by proposing that competences impose certain kinds of "demands" on the color visual system. An object re-identification competence demands that the experienced color of the object remain relatively *constant* in different perceptual situations. Other competences, such as figure-ground segregation, demand that the color of a particular object be experienced as sufficiently *different* from the colors of the neighboring regions. In many good cases of color perception, the demands of the relevant competences line up, and color vision can simultaneously fulfill its enhancement function with respect to all of them. We can call such cases "ideal." For example, seeing a ripe Red Delicious apple as red can simultaneously enhance figure-ground segregation, object (apple) recognition, property (ripeness) identification, and so on.[21]

[21] Perceivers can, of course, fail to manifest perceptual competences for various reasons. I might mistake a tomato for an apple, i.e., fail to exercise my object (tomato) identification competence,

In some other good cases, however, the demands of the relevant competences diverge and clash. We can call such cases "non-ideal." Non-ideal good cases differ from "bad" cases of color perception in that in the non-ideal cases color vision directly and non-accidentally enhances or enables *at least one* relevant competence, whereas in the bad cases color vision does not help the organism manifest *any* of the relevant competences.[22] Consider the example of negative afterimages. Suppose that you stare at a blue patch for a prolonged period, allowing your cones enough time to adapt to the stimulation and lose sensitivity. If you then turn your gaze to a white surface, a yellow patch appears. The neural processing that results in the experience of yellowness is not useful to you and does not help you manifest any relevant competences.[23] In non-ideal good cases, on the other hand, the perceptual circumstances are such that the color visual system is forced to "choose" between the conflicting demands of the relevant competences. This does not mean that the system fails, but that it fulfills its function *to the extent possible under those circumstances.* Recall the analogy of the cartographer who must choose what to represent in a map and how to represent it. Like the cartographer, the color visual system must sometimes choose how to best serve perception and action.[24,25]

---

even if my color visual system is doing everything right and the "demands" of the competences line up.

[22] The distinction between ideal and non-ideal good cases has nothing to do with the appeals that some philosophers make to "ideal conditions" in which the true colors of object surfaces are veridically perceived by normal perceivers. Under my view, color experiences are correct in both ideal and non-ideal cases, and the distinction is merely meant to capture the difference between perceptual situations where the demands of the relevant competences clash and situations where no such clash occurs.

[23] In the afterimage case the color visual system is engaged and a chromatic experience results. That said, we could plausibly extend the notion of color visual system failure to situations where the relevant kind of processing simply does not take place. In very low light, the visual perception of contours, objects, and scenes relies solely on achromatic rod vision, with no help from color vision. From the point of view of CE, any instance of visual perception in which color vision fails to be useful can be considered a "bad" case of color perception. This is because the role of color vision in the perceptual economy of the organism is understood purely in terms of enhancement. When such enhancement takes place, the color system fulfills its function, and we have a good case of color perception. When enhancement does not take place, we have a bad case of color perception. Just as a respiratory system can fail to fulfill its function due to some internal condition (e.g., pulmonary embolism) or some external condition (e.g., the presence of high levels of carbon monoxide in the air), the color visual system, too, can fail to fulfill its function when the light levels are too low.

[24] "Choices" should be understood loosely as the rules that the color system either follows or instantiates.

[25] I mentioned earlier (in note 9) that some readers might prefer to think of color vision as dividing into multiple systems with different aims. Those readers can now read me as suggesting that in the ideal cases the demands of the competences converge and the different systems can fulfill their functions simultaneously (by coming up with the same answer, so to speak). In non-ideal cases, on the other hand, one of the systems is forced to cede precedence to another. Many thanks to an anonymous reviewer for pointing this out.

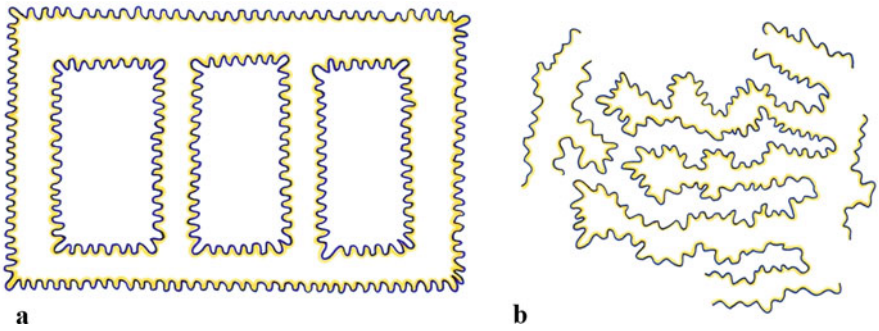


Fig. 1 A demonstration of "watercolor effect": the color of a yellow line flanking a darker contour of contrasting chromaticity appears to spread to cover the fully (**a**) or partially (**b**) enclosed area(s). Images are inspired by those found in Pinna et al. (2001)

4.2 Color Assimilation: Watercolor Effect

The perceived color of a target region sometimes shifts towards that of its neighbor and the chromatic contrast between the two regions is reduced (von Bezold 1874, 1876). This sort of assimilation of neighboring colors requires a particular kind of spatial organization; one way of inducing the effect is to intersperse small areas of color (Munker 1970). A particularly striking illustration of color assimilation is the so-called "watercolor effect," originally demonstrated by Pinna in 1987. In stimuli that produce this effect, adjacent regions are separated by two contiguous, differently colored boundary lines, one darker and the other lighter. The color of the lighter boundary then appears to spread to cover the enclosed area (Pinna et al. 2001, p. 2669). For example, in Fig. 1 (both **a** and **b**), the color of the yellow line flanking the darker blue contour spreads to cover the area defined by the darker boundary. As Pinna (2005) notes, two separate effects are visible in watercolor illusions: coloration effect and figural effect. The coloration effect is the apparent assimilative color spreading; the figural effect is the figure-ground organization in which the colored area is seen as a figure against a background.

Watercolor illusions are *illusory* in that there is not a three-dimensional figure where one is perceived; our experience imposes depth and volumetric impression onto a two-dimensional image. That said, illusory figural effects can provide clues as to how color vision subserves successful figure-ground segregation and object perception in more natural contexts. In this case it suggests that the visual system treats certain combinations of chromatic and achromatic edges as indicators of the presence of an object. In particular, it suggests that consistent edge color along a stretch of contour might be a figural cue that reliably correlates with objectness in our natural environments (von der Heydt and Pierson 2006, p. 337).[26]

[26] Pinna (2005) himself takes watercolor illusions to reveal a new principle of perceptual grouping and figure-ground segregation that he calls the "the asymmetric luminance contrast principle."

The figural effect in watercolor illusions is illusory, but how about the coloration effect? Does our color visual system fail when it presents some of the "white" areas in Fig. 1 as faint yellow? Many philosophers would answer this question in the positive, and argue that Fig. 1 tricks the system into producing an illusory experience. The problem is that if the watercolor effect reveals something about the way we normally compute and experience surface colors, then the color visual system failure might extend well into normal perception. More generally, if assimilation effects indicate color visual system failure, and if such effects are fairly commonplace, then much of our perceptual experience could turn out to be illusory.

One way to bring our perceptual experiences of Fig. 1 into the confines of correct color perception is to take the radically relativist route and argue that the enclosed areas in Fig. 1 really *are* yellow for perceivers like us in circumstances in which we perceive them as yellow, barring interference with the normal functioning of the color visual system (see McLaughlin 2003; Cohen 2009). In other words, if we relativize colors to particular perceivers and particular perceptual circumstances, we can maintain that the addition of certain types of bi-chromatic contours can sometimes literally change the color of a surface.

Although this strategy allows us to accommodate watercolor effects, it does not explain why we find these effects so puzzling. We seem naturally inclined to label watercolor effects illusory, and this inclination demands an explanation. It cannot just be that we are tricked into believing that the enclosed surface areas would produce these experiences in other contexts as well, e.g., in the absence of the appropriate sort of bi-chromatic contours.[27] We can know about the watercolor effect and not hold any such beliefs, and still have the intuition that something about our faint yellow experience is a bit off. So, what differentiates our experiences of watercolor illusions and our experiences of ripe Red Delicious apples? It seems that the relativist does not have much to offer here.

The perceptual pragmatist starts by asking if the experiences elicited by the images in Fig. 1 are useful to the perceiver. The enhancement of 3D figureness is not useful because it is illusory, but perhaps the coloration also helps the perceiver

According to this principle, ceteris paribus, given a boundary and an asymmetric luminance contrast on both sides of the boundary, the region with the less abrupt luminance gradient is perceived as the figure, and the region with the more abrupt luminance gradient is perceived as the background (ibid., p. 205). In Fig. 1, the luminance gradient is more abrupt where the dark contour directly meets the white background, and less abrupt where the dark contour meets the lighter contour which in turn meets the background. Although luminance contrast alone might be enough to bring about the figural effect, the coloration effect and the figural effects do seem to support and reinforce one another, as Pinna himself observes (see also von der Heydt and Pierson 2006, p. 334). In addition, Devinck et al. (2006) report that not only is a chromatic contrast between the two contours and between the contours and the background important for a *robust* watercolor effect, but that the effect can also be induced using equiluminant stimuli.

[27] Cohen explains the illusoriness of certain kinds of chromatic experiences by the non-veridicality of higher-level representations of the kind *surface x is orange to perceivers pretty much like me, in circumstances pretty much like those I normally encounter* (2007, p. 343). But we often find certain perceptual experiences puzzling when no such higher-level representations are involved.

discriminate some spatial properties that the stimuli do in fact instantiate. It might, for example, help the perceiver discern and make sense of the 2D spatial structure of the image by having some of the regions appear white and others yellowish. In **a**, the coloration might help us more efficiently perceive three jagged rectangle shapes confined within a larger jagged rectangle shape. In **b**, the enhancement effect might be even more profound, because the spatial structure is more complex.

Although this kind of usefulness is arguably rather minimal, it is usefulness nonetheless. Though much of our spatial perception "in the wild" involves the perception of 3D shapes, 2D patterns and shapes can also carry ecological significance, e.g., in the case of the surface patterns of animals and plants. In the language of CE, it seems at least plausible that the coloration enhances the manifestation of a 2D shape perception competence, a competence that makes it easier for perceivers to discern and interpret images, patterns, and scenes. And so there need not be anything illusory about the yellowish tint.

That said, the coloration does *not* enhance the manifestation of many other relevant perceptual competences. In fact, it seems to make it *more difficult* for perceivers to exercise competences that demand constancy and similarity of experience. For example, if we were to remove the yellow contours from **a**, a naïve observer might deny that she was observing the same surface, if none of the regions no longer appeared yellow to her. But we should not blame this on the color visual system. The system is still doing what it is supposed to be doing, as long as it is making it easier for the perceiver to manifest at least one relevant competence. That it cannot enhance the manifestation of *all* the relevant competences is a consequence of stimulus properties that pit the demands of the competences against one another.[28]

4.3 Simultaneous Color Contrast: Pink/Grey Petals

What looks yellow on an achromatic background might look greenish next to red and reddish next to green, and a normally red-appearing surface might look *even* redder next to green (Chevreul 1839, 1861). This phenomenon is known as "simultaneous color contrast," and psychologist Akiyoshi Kitaoka has created many powerful illustrations. Figure 2 depicts cartoonish flowering plants against two differently colored backgrounds and is inspired by some of Kitaoka's works.[29] Although the plants are physically identical in both frames, they appear grey against the white background (**b**) and pink against the cyan background (**a**), at least to most normal human color perceivers. In what follows, I will focus on the two types of

[28] Recall that textbook color illusions are usually designed to make certain kinds of perceptual effects more noticeable. In more ecologically realistic settings assimilation effects are likely to be less drastic, but also more useful.

[29] The RGB values of the background in Fig. 2 (**a**) are based on those used by Akiyoshi Kitaoka in the "Cherry blossom 2" section of his website.

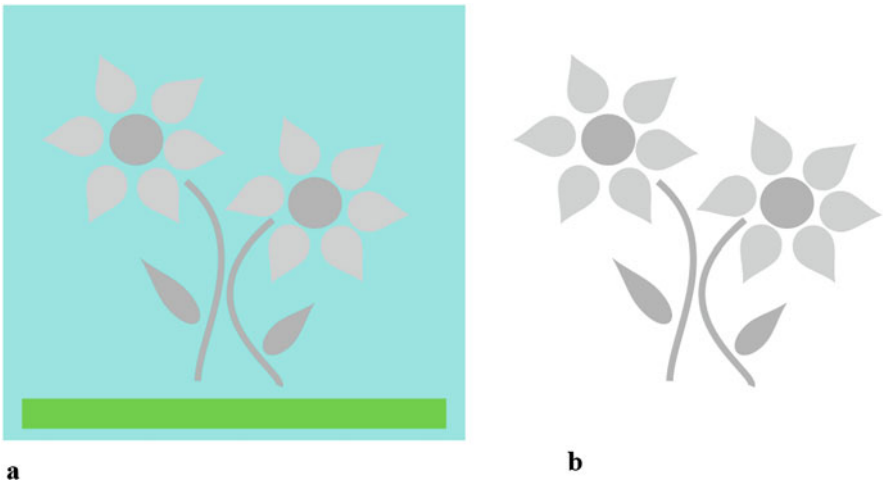


Fig. 2 A demonstration of simultaneous color contrast effect: physically identical flower shapes appear pink against cyan background (**a**) and grey against white background (**b**). Images are inspired by those found on Akiyoshi Kitaoka's website

chromatic experiences elicited by the "petal" regions in the two frames: pink and light grey. The challenge for the philosopher is to explain which (if either) of these chromatic experiences is correct, and why.

Many philosophers argue that each physically homogeneous region instantiates just one determinate, fine-grained color to a particular type of perceiver (e.g., Tye 2000; Byrne and Hilbert 2003). This entails that at most one type of chromatic experience of the petals can be correct (in that type of perceiver). The correctness question cannot be answered without privileging some specific perceptual conditions, and these include the color of the surround against which the target is viewed. It is doubtful that this specification can be done in a genuinely non-arbitrary manner.[30]

Although we can accommodate a great deal of perceptual variation by positing only *coarse-grained* stable colors (e.g., Hatfield 1992, 2003; see also Gert 2006), this move is unlikely to help here. Note that the motivation for views which attribute only coarse-grained colors to object surfaces (etc.) is often ecological and pragmatic: chromatic experiences can help perceivers identify objects and properties as long as those experiences stay within a certain range, e.g., a ripe apple might appear one shade of red in daylight and another shade of red under incandescent lighting, and this is enough to help us recognize it as the same apple. An exact hue match is not required for such identification, but some degree of constancy is necessary. If the perceived color of the apple shifts from red to green, for example, the chromatic experience is no longer useful in this sense. The case with the petals seems

[30]This point has been forcefully argued by Hardin (1993) and Cohen (2009).

analogous; our experiences simply do not fall within a range that would be conducive to such identification. This in turn suggests that the grey and pink experiences cannot both be correct, and that we again need some non-arbitrary way of privileging one of these perceptual variants over the other.[31]

Philosophers who attribute stable (fine-grained or coarse-grained) colors to surfaces therefore all seem to be in the same boat; they have to say that at least some instances of simultaneous color contrast constitute color visual system failure.[32] Letting go of the notion of stable color can help defend the color visual system against this charge, and one way to do this is to maintain that the petals really *are* grey when specific perceivers observe them against a white surround, just like they really *are* pink when the same perceivers observe them against a cyan background. But once again this sort of relativism seems ill-equipped to deal with the task of explaining the difference between textbook color illusions and more straightforward cases of correct color perception.

The perceptual pragmatist can make progress here by explaining the shifts in the perceived hue of the petals in terms of what is useful to the perceiver: seeing the petals as pink against the cyan background increases the contrast between the petals and the background, making the petals pop out. Although it is difficult to imagine what it would be like to experience the petals as achromatic grey against the cyan background, it seems reasonable to assume that the grey would more easily blend with the cyan, making it more difficult to segment and interpret the image. But this is just a beginning – we still need an account of the difference between seeing the petals as pink and seeing a ripe apple as red. If both experiences are useful, then why does it seem so natural to label one experience illusory and the other perfectly correct?

CE can help fill in the gaps. Observing Fig. 2 constitutes a non-ideal perceptual case where the relevant competences place divergent demands on the color visual system, whereas in the ripe apple case the demands of the relevant competences converge.[33] To understand how the demands diverge in the petals case, let us start by

[31] Hatfield (2003) suggests that the true coarse-grained colors that objects instantiate can be veridically perceived by species-specific normal perceivers in ecologically standard conditions. But "ecologically standard conditions" cannot just refer to ecologically standard *lighting* conditions (e.g., daylight for humans). If it did, then we could not determine which color the petals in Fig. 2 *really* instantiate and we also could not evaluate the correctness of the color experiences elicited by the image. On the other hand, specifying ecologically standard surround colors seems like a difficult task. Natural scenes instantiate a wide variety of different colors, including whites and brilliant cyans, and it seems arbitrary to rule some of them out as being ecologically non-standard.

[32] Views that posit stable fine-grained colors entail widespread color visual system failure (because hue shifts at the level of fine-grained colors are extremely common), whereas views that posit only coarse-grained colors can restrict such failure to special cases. But the latter views need to explain why perceptual variation due to simultaneous contrast is acceptable when it occurs within color types, but unacceptable when it crosses those types. This is particularly difficult if it turns out that color type-crossing perceptual variation is *useful* to the perceiving organism, since considerations of usefulness are often what often motivate such views to begin with.

[33] Imagine a human color normal viewing a ripe Red Delicious against green foliage. Figure-ground segregation and object perception competences "demand" that the apple be perceived as

supposing that a normal human color perceiver is observing a printout of Fig. 2. Since she is observing a picture rather than a natural scene, the list of relevant perceptual competences is different from the apple case. There are no actual petals to be detected, identified, or re-identified, just a flat image *depicting* petals against backgrounds of different chromaticities. Yet there are still relevant competences placing demands on the color visual system here. For example, the 2D shape perception competence demands that the regions with different physical properties be perceived as sufficiently different in hue, saturation and/or lightness. Where the petals are depicted against the white background in **b**, the achromatic luminance contrast alone seems to make them pop out (simultaneous luminance contrast might play some role here). In **a**, the luminance contrast is softer, and there is more of a danger of perceptual blending of the different regions. But in addition to luminance contrast, we now have chromatic contrast between the regions. By enhancing this contrast (and giving rise to an experience of pinkness), the color visual system can help ensure that the image is accurately segmented.

But once again this enhancement of the shape perception competence occurs at the expense of some other competences. If we secretly cut out a petal from **a**, hand it to our subject, and ask her if it is one of the petals she previously observed against the cyan background, she might well say no. This is because the petal no longer appears pink to her when removed from its original context. That is, seeing the petal as pink against the cyan background helps our subject manifest her shape perception competence, while simultaneously hindering the manifestation of her re-identification competence.

CE entails that the apple case and the petal case are both instances of successful, correct color perception. It also entails that there is a clear difference between the two: in the apple case the color visual system can fulfill its competence-enhancement function more fully because the demands of the relevant competences line up; in the petal case the system cannot simultaneously enhance the manifestation of all the relevant competences but has to "choose" between them. This sort of conflict is a relatively rare occurrence, which explains the strangeness of the resulting experience.

---

chromatically different from the surrounding foliage. Seeing the apple as red helps achieve this, although seeing the apple as some other color sufficiently different from the perceived color of the foliage would also suffice. On the other hand, the color visual system can only help the person manifest her ripe Red Delicious apple identification competence by outputting an experience sufficiently similar to the experiences *usually* elicited by ripe Red Delicious apples. An object re-identification competence "demands" that the perceived color of the apple does not vary too much with context, i.e., that the apple does not look dramatically different in different lighting conditions or against different backgrounds. When I say that the demands of these competences converge, I simply mean that a (normal) color visual system *can* simultaneously help an organism manifest all these competences.

5 Objections, Replies, and Further Developments

I have proposed that we understand color vision as being competence-embedded and shown how the framework of competence-embeddedness can shed light on some puzzling color phenomena. In this section, I will consider some pertinent objections and suggest ways in which the view could be developed further.

5.1 Intuitions and Common Sense

It is common to criticize philosophical views of color and color perception for being counterintuitive or in conflict with common sense. Supposedly, when we experience colors, we experience them as being stable (perhaps even mind-independent) properties of objects. And, supposedly, the way we naturally think and talk about color suggests that there can be genuine disagreement about the colors of things. When a theory conflicts with these notions, it is often taken as evidence of the theory's falsehood. Tye's (2012) critique of Cohen's (2007, 2009) relationalism is a good example. Tye considers a hypothetical scenario where a perceiver observing a ball reports "That ball is yellow" and another perceiver observing the same ball reports "That ball is not yellow; it's pink." For Tye, the color attributions in this case are in direct competition. Intuitively, suggests Tye, there is genuine disagreement between the two perceivers, and at best only one of them can be right about the color of the ball. Because Cohen's relationalism entails that the ball instantiates both yellowness and pinkness in this scenario, and because this verdict is inconsistent with the alleged deliverances of intuition, Tye concludes that Cohen's view is wrong. Since CE also entails that both the yellow-involving experience and the pink-involving experience are simultaneously correct (or at least *can* be), the same objection applies.

There are at least two ways to respond to Tye. First, we can deny his claim about what is intuitively true about color and color perception. Second, we can resist the view that *commonsense* intuitions and *ordinary* discourse are what settle matters in philosophy of color. I shall focus on the first response here.

Here is my reading of Tye's hypothetical scenario: it is true that if I saw a ball as yellow and somebody else claimed to see it as pink, I might be somewhat worried. But, crucially, I would not worry because I did not find it metaphysically possible for the ball to simultaneously instantiate yellowness and pinkness. I would worry if I took my fellow perceiver to be a normal human color perceiver making a claim about the way the ball chromatically appears to normal human color perceivers in the conditions that we were both in at the time (this, I take it, is how we use color talk in most ordinary situations). Despite rampant interpersonal variation in color perception,[34] two normal perceivers of the same species do not usually see the same object as yellow and pink in the same conditions. Thus, if my fellow observer saw the ball

[34] See, e.g., Schefrin and Werner (1990), Kuehni (2004).

as pink, I might suspect that there was something abnormal going on with her perceptual processing; perhaps she was intoxicated or experiencing a migraine headache with visual aura. But if she then explained that she was, in fact, an atypical human color perceiver (with mutant cones or "an usual physiological condition" as Tye stipulates in his original example), or a normal *alien* color perceiver, my worry would dissipate.[35] In the language of CE, as long as she was manifesting some relevant competence (human or alien) by perceiving the ball as being pink, all would be fine and good.[36]

It is therefore not at all clear that intuitions are on Tye's side. This becomes even clearer when we consider the perception of fine-grained hues. Contrary to Tye, I have little trouble accepting that what is one shade of red to me might be a different shade of red to you, and that objects might simultaneously instantiate a variety of fine-grained hues.[37] Tye does not think that perceptual variation is a good enough to reason to let go of the notion of stable fine-grained object color. He denies that our inability to non-arbitrarily assign such colors to objects means that no such colors exist. This might be true, but Tye still needs to provide an argument for his claim that stable fine-grained colors do in fact exist. As far as I can see, his only reason for accepting this claim is common sense dogmatism. In his own words, "in the absence of a good reason to disbelieve the ordinary view, we are warranted in believing that view" (2012, p. 299). But Tye has neither shown that the view that he is putting forward as the commonsense view really *is* the commonsense view (and the only commonsense view), nor has he shown that the commonsense view should hold the kind of power that he claims it does (i.e., that dogmatism is the way to go when it comes to color). In addition, there are, I believe, good reasons to disbelieve Tye's view. It entails rampant color misperception and renders color facts unknowable, among other things.

An attentive reader might now point out that the three desiderata sketched in Sect. 2 are also partially motivated by intuitions about color perception. Are not the desiderata therefore just as suspect as the intuitions on which Tye bases his criticism of Cohen? I think not. The intuitions I appeal to are philosophical intuitions that are much more robust than Tye's commonsense intuitions about the stability of true color. Although Tye's yellow/pink example might elicit in some people the kind of responses he is after, this is much less likely to happen if more intricate hypothetical scenarios are used. The three desiderata, on the other hand, are based on intuitions that are widely shared among philosophers who work on color and that do not rely on the use of specific thought experiments or scenarios.

[35] This way of thinking comes naturally to me. Empirical work by Cohen and Nichols (2010) suggest that it comes naturally to others as well.

[36] Note that CE also entails that many/most of the color experiences of atypical human color perceivers (e.g., dichromats or anomalous trichromats) are correct.

[37] That said, I acknowledge that we can disagree about the referents of color terms. For example, if an object looks distinctly mint-green to me and a friend claims that it is turquoise, I might take her to be confused about the referents of "mint-green" and "turquoise," on the assumption that the two of us live in fairly similar phenomenal worlds.

5.2 The Second Desideratum and Normal Illusion Talk

Another worry is that CE does not satisfy the second desideratum and cannot accommodate normal illusion talk. To elaborate on this worry, let us imagine a hallucinogen that interferes with the stages of cortical processing that correlate with conscious color experience, causing users to project seemingly random colors onto objects with which they perceptually interact. Let us dictate that all the contrast processing in the brain functions normally and the projecting of colors respects this processing in the sense that identical colors are never projected onto adjacent regions separated by a contour (this is a sense in which the projection is *not* random). If scenes are appropriately segmented and if the color visual system contributes to the segmentation, then it looks like at least one perceptual competence is manifested with the help of the color visual system. CE therefore entails that the distorted color experiences are correct. Yet, at the same time, this sounds like a fairly straightforward case of illusory color perception, since the experienced colors have little to do with the properties of the surfaces in question.

Could we rule this case out on the basis of the color experience being abnormally caused? This might seem plausible at first. The color visual system is not functioning the way it normally functions; the hallucinogen interferes with the usual neural pathways and distorts the chromatic experience. So perhaps a way to defuse this objection is to simply require that color perception be competence-enhancing *in the normal way*. This requirement would also conveniently rule out experiences that occur due to some focal dysfunction in the brain, e.g., a migraine with visual aura, a head trauma, or the influence of many (ordinary) hallucinogens. In all such cases the color experience is *potentially* useful, but only *accidentally* so. A translucent tomato-shaped red visual aura covering the part of a visual field where an actual tomato is located might accidentally help a perceiver identify a tomato as a tomato, but it is not the aim of the visual system to enhance this competence by producing the aura.

Notice, however, that in our imaginary hallucinogen case the color experiences are not accidentally useful. They are useful because the processing of chromatic edges plays a part in how the perceiver experiences the chromatic properties of her environment. Because the drug only interferes with the later stages of color perception and respects the competence-enhancing earlier processing, there is a sense in which the color visual system has *already fulfilled its function*. So perhaps the proponent of CE should just bite the bullet and accept that literally any color experience that directly and non-accidentally subserves the relevant competences is correct, even if there is something unusual about the experience itself.

Although I am sympathetic to this option, there is no denying that the colors experienced under the influence of the imaginary hallucinogen could be immensely confusing. A tomato might look chartreuse, trees purple, and the sky maroon. It might therefore seem counterintuitive to claim that such experiences are perfectly useful and correct. But note that an experience could be *minimally* useful/correct and confusing at the same time. To explain what I mean by this, I will now turn to the third objection.

5.3 When the Chromatic Experiences of Normal Perceivers Dramatically Diverge: "The Dress," Etc

The last objection borrows its motivation from cases where the color experiences of normal human color perceivers radically diverge, and where the divergence is not connected to differences in the absorption spectra of retinal cones, to normal neural adaptation, or to migraine headaches, hallucinogens, or any other unusual state of affairs. Perhaps the most striking example is the controversy surrounding "the dress," i.e., the overexposed photograph of a dress that first became a viral internet sensation in 2015. The illumination cues in the image are ambiguous, and perceivers experience the colors of the dress in vastly different ways. Some see the dress as white and gold, others as black and blue, and a small minority as some other combination of colors.[38]

"The dress" poses a problem for CE because CE appears to entail that the perceivers in the different camps all experience the colors in the image correctly, as long as those experiences are tied to the manifestation of some relevant competence. But this sounds rather strange, especially if we think that this sort of divergence could potentially extend to situations where observers are viewing actual objects (and not just photographs of objects) in some atypical viewing conditions. To see why, imagine that two subjects are viewing the actual dress through an aperture in a laboratory where vision scientists have carefully created ambiguous lighting cues comparable to those in the original photograph (the practical possibility of this need not concern us here). Now imagine that Subject A sees the dress as black and blue and Subject B sees it as white and gold, but when they are shown the same dress in normal daylight conditions, they both see it as black and blue. The question is: is it reasonable to insist that the subjects' divergent color-involving experiences in the aperture viewing condition are both correct?

There are at least two plausible ways the proponent of CE could respond. First, she could argue that the two experiences are equally correct and that the controversy concerning the color of dress is purely doxastic.[39] In short, Subject B's experience of the dress as white and gold is not itself mistaken but if she then forms the belief, based on that experience, that the dress looks white and gold to her (and to perceivers like her) in most ordinary perceptual circumstances, she is clearly wrong. And thus the *belief* is erroneous, not the experience itself. That said, it is unclear whether this strategy can appease intuitions about the two experiences themselves being on an unequal footing. Many would likely argue that there is still something askew about Subject B's white-and-gold experience even in the absence of any subsequent beliefs about the way the dress normally appears.

[38] The differences might reflect certain "internal priors" of the perceivers: the perceivers in the different camps might "favor" and expect different kinds of illuminants (see Lafer-Sousa et al. 2015).

[39] This option is inspired by Cohen (2009).

Luckily, CE also admits of a more complex response. Note that the experience of the dress as black and blue in the aperture viewing condition might help Subject A manifest an additional competence: *object re-identification*. Because her chromatic experience remains relatively stable across the two conditions, it is easier for her to recognize the dress as being the same dress (or at least a dress with similar material properties). And if her experience helps her manifest *more* competences, then perhaps her experience is also *more* correct than Subject B's experience in the aperture condition.

Although this might sound a bit odd initially, recall that CE analyzes correctness in terms of usefulness. Since usefulness admits of degrees (both a key and a hammer can be useful for getting through a locked door, but the right kind of key is usually more useful than a hammer), then arguably correctness-as-usefulness can too. It is useful (competence-enhancing) to perceive the dress as white and gold, but even more useful (competence-enhancing) to perceive it as black and blue. Therefore, we could say that *it is also more correct to perceive the dress as black and blue*.[40]

Going back to the earlier example, we could now say that an experience of a visual scene under the influence of the imaginary hallucinogen is minimally correct, but less correct than an experience of the same scene when the color visual system is functioning normally. This would allow us to analyze the apparent illusoriness of some perceptual cases as having to do with a lesser degree of correctness than we might normally expect, and the disagreement between some perceptual variants as a disagreement about which variant is *more* correct. Notice, however, that this sort of analysis does not apply to cases like the pink/grey petals where the demands of the relevant competences themselves are in conflict. The structural dissimilarity between the hallucinogen case and the pink/grey petals case is reflected in the structural dissimilarity of the respective explanations.

6 Conclusion: The Three Desiderata Revisited

CE appears to largely avoid the pitfalls that plague many other philosophical views of color perception. First, it does not entail widespread color visual system failure: insofar as most of our color experiences result from processing that directly and non-accidentally subserves the manifestation of relevant competences, the view entails that color visual systems usually function the way they are supposed to function (*desideratum 1*). At the same time, CE does allow for instances of unsuccessful color perception: if a particular color experience results from processing that does not directly and non-accidentally contribute to the manifestation of any relevant

[40] "Non-pragmatist" views seem antithetical to this line of reasoning. For example, if color is equated with some microphysical property and veridical color perception is understood in terms of accurate detection of this property, then there exists a rigid binary of correct and incorrect color perception. CE is consistent with both the rigid boundary option and the gradated option, and those who find the idea of degrees of correctness unsavory can still accept the core tenets of CE.

competences, it is incorrect (*desideratum 2).* Finally, CE allows us to unambiguously differentiate between instances of correct color perception and instances of incorrect color perception, and to use these standards to evaluate and explain a wide variety of color perceptual phenomena (*desideratum 3*). A great deal of explanatory power comes from the distinction between ideal and non-ideal good cases of color perception. Many textbook color illusions count as non-ideal good cases, and CE allows us to explain their apparent strangeness without attributing failure to the system itself. In addition, because CE allows for degrees of correctness, it allows us to maintain that some color experiences are more correct than others, even if all the candidate experiences result from competence-enhancing processing.

Acknowledgements I would like to thank Gary Hatfield, Lisa Miracchi Titus, Elizabeth Johnson, Quayshawn Spencer, Penelope Maddy, Evan Sommers, Charles Leitz, Jeffrey Schatz, Eugene Vaynberg, Youngbin Yoon, Kate Nicole Hoffman, and two anonymous reviewers for their very helpful comments and suggestions.

References

Akins K (2001) More than mere colouring: a dialog between philosophy and neuroscience on the nature of spectral vision. In: Fitzpatrick SM, Bruer JT (eds) Carving our destiny: scientific research faces a new millennium. Joseph Henry Press, Washington, DC, pp 77–116

Akins KA, Hahn M (2014) More than mere colouring: the role of spectral information in human vision. Br J Philos Sci 65:125–171

Allen K (2016) A naïve realist theory of colour. Oxford University Press, Oxford

Baxandall M (1988) Painting and experience in fifteenth-century Italy: a primer in the social history of pictorial style. Oxford University Press, Oxford

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147

Boghossian P, Velleman JD (1989) Colour as a secondary quality. Mind 98:81–103

Byrne A, Hilbert D (1997) Colors and reflectances. In: Byrne A, Hilbert D (eds) Readings on color, I: the philosophy of color. MIT Press, Cambridge, pp 262–288

Byrne A, Hilbert DR (2003) Color realism and color science. Behav Brain Sci 26:3–64

Campbell J (1993) A simple view of color. In: Haldane JJ, Wright C (eds) Reality: representation and projection. Oxford University Press, Oxford, pp 257–268

Chevreul ME (1839) De la loi du contraste simultané des couleurs et de l'assortiment des objets colorés. Pitois-Levrault, Paris

Chevreul ME (1861) The laws of contrast of color: and their application to the arts (Trans: Spanton J). Routledge, Warne; London

Chirimuuta M (2015) Outside color: perceptual science and the puzzle of color in philosophy. MIT Press, Cambridge

Chittka L, Briscoe A (2001) Why sensory ecology needs to become more evolutionary – insect color vision as a case in point. In: Barth FG, Schmid A (eds) Ecology of sensing. Springer-Verlag, Berlin, pp 19–37

Cohen J (2003) Perceptual variation, realism, and relativization, or: how I learned to stop worrying and love variations in color vision. Behav Brain Sci 26:25–26

Cohen J (2007) A relationalist's guide to error about color perception. Noûs 41:335–353

Cohen J (2009) The red & the Real: an essay on color ontology. Oxford University Press, Oxford

Cohen J, Nichols S (2010) Colours, colour relationalism and the deliverances of introspection. Analysis 70:218–228
Cummins R (1975) Functional analysis. J Philos 72:741–765
De Valois KK, Switkes E (1983) Simultaneous masking between chromatic and luminance gratings. J Opt Soc Am 72:11–18
Devinck F, Hardy JL, Delahunt PB, Spillman L, Werner JS (2006) Illusory spreading of watercolor. J Vis 6:625–633. https://doi.org/10.1167/6.5.7
Dominy NJ, Lucas PW (2001) Ecological importance of trichromatic vision to primates. Nature 410:363–366. https://doi.org/10.1038/35066567
Dresp-Langley B, Reeves A (2014) Color and figure-ground: from signals to qualia. In: Geremek A, Greenlee M, Magnussen S (eds) Perception beyond gestalt: progress in vision research. Psychology Press, New York
Garg AK, Li P, Rashid MS, Callaway EM (2019) Color and orientation are jointly coded and spatially organized in primate primary visual cortex. Science 364:1275–1279. https://doi.org/10.1126/science.aaw5868
Gegenfurtner KR, Rieger J (2000) Sensory and cognitive contributions of color to the recognition of natural scenes. Curr Biol 10:805–808. https://doi.org/10.1016/S0960-9822(00)00563-7
Gert J (2006) A realistic colour realism. Australas J Philos 84:565–589
Gert J (2018) Primitive colors: a case study in neo-pragmatist metaphysics and philosophy of perception. Oxford University Press, Oxford
Godfrey-Smith P (1994) A modern history theory of functions. Nôus 28:344–362
Greco J (2007) The nature of ability and the purpose of knowledge. Philos Issues 17:57–69
Hansen T, Gegenfurtner KR (2017) Color contributes to object-contour perception in natural scenes. J Vis 17. https://doi.org/10.1167/17.3.14
Hardcastle VG (2002) On the normativity of functions. In: Ariew A, Cummins R, Perlman M (eds) Functions: new essays in the philosophy of psychology and biology. Oxford University Press, Oxford, pp 144–156
Hardin CL (1993) Color for philosophers: unweaving the rainbow. Hackett Publishing, Indianapolis
Hatfield G (1992) Color perception and neural encoding: does metameric matching entail a loss of information. Proc Biennial Meeting Philos Sci Assoc 1:492–504
Hatfield G (2003) Objectivity and subjectivity revisited: color as a psychobiological property. In: Mausfeld R, Heyer D (eds) Colour perception: mind and the physical world. Oxford University Press, Oxford, pp 187–202
von der Heydt R, Pierson R (2006) Dissociation of color and figure-ground effects in the watercolor illusion. Spat Vis 19:323–340
Hilbert D (1992) What is color vision? Philos Stud 68:351–370
Hilbert DR (1987) Color and color perception: a study in anthropocentric realism. Stanford University CSLI, Stanford
Hornett W (2021) Perceptual capacities, discrimination, senses. Synthese. https://doi.org/10.1007/s11229-021-03410-2
Jacobs GH (1981) The distribution and nature of colour vision among the mammals. Biol Rev 68:413–471
Kingdom F (2008) Perceiving light versus material. Vis Res 48:2090–2105. https://doi.org/10.1016/j.visres.2008.03.020
Kitaoka A. *Akiyoshi's illusion pages*. http://www.ritsumei.ac.jp/~akitaoka/index-e.html. Accessed 28 Nov 2021
Kuehni RG (2004) Variability in unique hue selection: a surprising phenomenon. Color Res Appl 29:158–162. https://doi.org/10.1002/col.10237
Lafer-Sousa R, Hermann KL, Conway BR (2015) Striking individual differences in color perception uncovered by 'the dress' photograph. Curr Biol 25:R523–R548. https://doi.org/10.1016/j.cub.2015.04.053

Levin J (2000) Dispositional theories of color and the claims of common sense. Philos Stud 100: 151–174
Matthen M (2005) Seeing, doing, and knowing: a philosophical theory of sense perception. Oxford University Press, Oxford
Maund JB (2006) The illusory theory of colour: an anti-realist theory. Dialectica 60:254–268
McLaughlin B (2003) The place of color in nature. In: Mausfeld R, Heyer D (eds) Colour perception: mind and the physical world. Oxford University Press, Oxford, pp 254–268
Mendelovici A (2016) Why tracking theories should allow for clean cases of reliable misrepresentation. Disputatio 8:57–92
Miracchi L (2015) Competence to know. Philos Stud 172:29–56
Miracchi L (2017) Perception first. J Philos 114:629–677
Mollon JD (1989) 'Tho' she kneel'd in that place where they grew.' The uses and origins of primate colour vision. J Exp Biol 146:21–38
Moutoussis K (2015) The physiology and psychophysics of the color-form relationship: a review. Front Psychol 6:Article 1407. https://doi.org/10.3389/fpsyg.2015.01407
Munker H (1970) Farbige Gitter, Abbildung auf der Netzhaut und übertragungstheoretische Beschreibung der Farbwahrnehmung. Habilitationsschrift, Ludwig-Maximilians-Universität, München
Neander K (2017) A mark of the mental: in defense of informational teleosemantics. MIT Press, Cambridge
Palmer SE, Brooks JL (2008) Edge-region grouping in figure-ground organization and depth perception. J Exp Psychol Hum Percept Perform 34:1353–1371. https://doi.org/10.1037/a0012729
Paramei GV, van Leeuwen C (2016) Editorial: color and form perception: straddling the boundary. Front Psychol 7:104. https://doi.org/10.3389/fpsyg.2016.00104
Peterson M (2015) Low-level and high-level contributions to figure-ground organization. In: Wagemans J (ed) The Oxford handbook of perceptual organization. Oxford University Press, Oxford, pp 259–280
Pinna B (1987) Un effetto di colorazione. In: Majer V, Maeran M, Santinello M (eds) Il laboratorio e la città. XXI Congresso degli Psicologi Italiani. Edizioni SIPs. Società Italiana di Psicologia, Milano, p 158
Pinna B (2005) The role of the Gestalt principle of similarity in the watercolor illusion. Spat Vis 18: 185–207
Pinna B, Brelstaff G, Spillmann L (2001) Surface color from boundaries: a new 'watercolor' illusion. Vis Res 41:2669–2676
Rogers-Ramachandran DC, Ramachandran VS (1998) Psychophysical evidence for boundary and surface systems in human vision. Vis Res 38:71–77
Schefrin B, Werner J (1990) Loci of spectral unique hues throughout the lifespan. J Opt Soc Am A 7:305–311
Schellenberg S (2018) The unity of perception: content, consciousness, evidence. Oxford University Press, Oxford
Schwartz PH (2002) The continuing usefulness account of proper function. In: Ariew A, Cummins R, Perlman M (eds) Functions: new essays in the philosophy of psychology and biology. Oxford University Press, Oxford, pp 244–260
Shapley R, Hawken M, Johnson E (2014) Color in the primary visual cortex. In: Werner JS, Chalupa LM (eds) The new visual neurosciences. MIT Press, Cambridge, pp 567–586
Shevell S, Kingdom F (2008) Color in complex scenes. Annu Rev Psychol 59:143–166. https://doi.org/10.1146/annurev.psych.59.103006.093619
Smithson HE (2015) Perceptual organization of color. In: Wagemans J (ed) The Oxford handbook of perceptual organization. Oxford University Press, Oxford, pp 436–465
Sosa E (2007) A virtue epistemology: apt belief and reflective knowledge, vol 1. Oxford University Press, Oxford
Sosa E (2015) Judgment and agency. Oxford University Press, Oxford

Tanaka J, Weiskopf D, Williams P (2001) The role of color in high-level vision. Trends Cogn Sci 5: 211. https://doi.org/10.1016/S1364-6613(00)01626-0
Thompson E (1995) Colour vision: a study in cognitive science and the philosophy of perception. Routledge, London
Troscianko T, Montagnon R, Le Clerc J, Malbert E, Chanteau P-L (1991) The role of colour as a monocular depth cue. Vis Res 31:1923–1930
Tye M (2000) Consciousness, color and content. MIT Press, Cambridge
Tye M (2012) Cohen on color relationalism. analytic. Philosophy 53:297–305
von Bezold W (1874) Die Farbenlehre im Hinblick auf Kunst und Kunstgewerbe. G. Westermann, Braunschweig
von Bezold W (1876) The theory of color and its relation to art and art-industry (Trans: Koehler SR). Prang and Company
Wright L (1973) Functions. Philoso Rev 82:139–168
Wright L (1976) Teleological explanations: an etiological analysis of goals and functions. University of California Press, Berkeley
Yamagishi N, Melara RD (2001) Informational primacy of visual dimensions: specialized roles for luminance and chromaticity in figure-ground perception. Percept Psychophys 63:824–826

Menstrual Cycles as Key to Embodied Synchronisation

Ainhoa Rodriguez-Muguruza

Abstract The concept of embodiment has allowed philosophical discussions to refer to cognition as the result of a series of interactions occurring both within and outside the organism. The factors that organisms look at when keeping up with that synchrony are defined as timekeepers, elements that set the pace of specific rhythms. In this paper, I elaborate on the hypothesis that, if an all-encompassing sense of physiological and psychosocial synchrony is essential for the enablement of the organism's cognitive capacity and, ultimately, overall well-being, the rhythm menstrual cycles perform must be considered as a key part of the process of attunement through which menstruating organisms understand reality and access activities of sense-making. Following the belief that lack of academic, scientific, and social knowledge on menstrual cycles has kept menstruating organisms from successful synchronisation and, consequently, the correct articulation of their cognition, I postulate throughout this paper menstrual cycles as a crucial tool for the resynchronisation of menstruating organisms with their physiological and psychological environment.

Keywords Menstrual cycle · Participatory sense-making · Cognitive embodiment · Enactivism · Cognitive enablement

1 Introduction

The concept of embodiment has allowed philosophical discussions to refer to cognition as the result of a series of interactions occurring both within and outside the organism. The factors that organisms look at when keeping up with that synchrony are defined as timekeepers, elements that set the pace of specific rhythms. In this paper, I elaborate on the hypothesis that, if an all-encompassing sense of physiological and psychosocial synchrony is essential for the enablement of the

A. Rodriguez-Muguruza (✉)
Department of Logic and Philosophy of Science, University of the Basque Country, Donostia – San Sebastián, Spain
e-mail: arodriguez974@ikasle.ehu.eus

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_6

organism's cognitive capacity and, ultimately, overall well-being, the rhythm menstrual cycles perform must be considered as a key part of the process of attunement through which menstruating organisms[1] understand reality and access activities of sense-making. Following the belief that lack of academic, scientific, and social knowledge on menstrual cycles has kept menstruating organisms from successful synchronisation and, consequently, the correct articulation of their cognition, I develop throughout this paper the arguments that postulate menstrual cycles as a crucial tool for the resynchronisation of menstruating organisms with their physiological and psychological environment.

The purpose of this paper endorses a definition of cognition based on an organism-environment connection that is inherently ascribed to a rhythm: the menstrual rhythm. In Sect. 1, I present an overview of the philosophical discussions on cognition and I underline the integration of the concepts of synchronisation, coupling, and participatory sense-making; terms that will prove crucial for the definition of the capacity of an organism to articulate cognition in coordination with environmental, physiological, and social interactions. The analysis of the vocabulary produced in the philosophical discussions on embodied cognition provides the paper with the tools needed to extend the conversation about the synchronisation of organisms with their body-rhythms to rhythms not taken into consideration until now. As such, Sect. 1 also delves into the cognitive consequences of both total and partial failure to synchronise with a particular body-rhythm.

In Sect. 2, I aim to provide support for the philosophical intuitions presented and elaborated in Sect. 1 on the role that synchronisation with body-rhythms has in the well-being of the organism. Through an outline of the research within the area of chronobiology, in Sect. 2, I draw a general picture of the role of body-rhythms, with a more detailed explanation of the effects body-rhythms have in cognition and focusing on the hormonal interactions behind said rhythms. Throughout the section, I underline the emphasis that chronobiology has put on circadian rhythms, devoting the second part of the section to an overview of the very limited literature available on other rhythms. This overview postulates menstrual cycles as an example of an infradian rhythm and provides a summary of the hormonal phenomena involved in menstrual cyclicity, introducing the question of whether menstrual rhythms could also have cognitive effects and enquiring over the possibility of whether that lack of understanding on menstrual cognitive changes has hindered the well-being of menstruating organisms.

With the purpose of answering both these questions, in Sect. 3, I further discuss the cognitive changes connected to circadian rhythms and how societal norms and activities have been developed to accommodate and optimise those changes. In Sect. 3, I provide examples of psychosocial practices that have been standardised to

[1] Throughout the paper, I will refer to the organisms subject to synchronisation through the menstrual cycle as 'menstruating organisms', with this phrase applying to all human bodies that experience natural menstrual cycles, regardless of their gender identity. Organisms experiencing natural menstrual cycles do not include users of hormonal birth control, hormone-replacement therapy, or in-vitro fertilisation, since their hormonal profile is altered by exogenous factors.

support periods of cognitive peak capacities throughout the day, suggesting that certain societal practices could invite and favour the synchronisation of the organism to circadian rhythms. Elaborating on the cognitive effects derived from menstrual cycles, in this section, I point out the lack of social recognition of menstrual rhythms and consequent absence of any shared practices that consider the accommodation of menstrual cognitive variations. To account for the positive effect adopting practices that recognise menstrual rhythmicity can have in menstruating organisms, Sect. 3 appraises research on cognitive improvements observed in menstruating athletes after training cadence and intensity were adapted to their menstrual cycles.

To elaborate on the procedure necessary to widen the awareness society has over menstrual cycles, in Sect. 4, I discuss whether the lack of social consideration for menstrual rhythms, together with the limited chronobiological literature available on non-circadian rhythms have kept menstruating organisms from perceiving themselves as rhythmically attuned and, consequently, have deteriorated their capacity to synchronise with those rhythms. In this section, I reclaim the knowledge on the menstrual cycle as key for the resynchronisation of menstruating organisms and underline it as an indispensable factor for menstruating organisms to successfully engage in participatory-sense making.

This paper is the first philosophical piece applying concepts on synchronisation, entrainment, and participatory sense-making to non-circadian rhythms and, most accurately, menstrual rhythms. It is also the first paper postulating menstrual cycles as an essential element of the cognition of menstruating organisms, reclaiming the philosophical relevance of this body process.

2 The Evolution of Synchronised Cognition

To explore the role the menstrual cycle might have in the cognitive enablement of an organism, it is essential to first provide an overview of the philosophical claims that have advocated for the need for the body to be included back into discussions of the mind. Research on the idea of embodiment has, in fact, enabled philosophy to understand cognition as a result of interactions between the physiology of an organism, their cognitive capacities, and the environment.

2.1 A Coordinated Definition of Cognition

Varela defined cognition as an intertwined network of interactions embedded in a multi-leveled manifold (Thompson and Varela 2001), distinguishing between endogenous processes within the organism, and exogenous responses involving the environment. Varela's paradigm of cognition is key in the incorporation of bodily rhythms into the discussions on cognition, resulting in an innovative understanding of how the mind ought to be addressed and the elements that could be

included when discussing cognitive capabilities. Bringing the mind, the body, and the environment together under the same system, Varela referred to the network within which they are embedded as 'cycles of operation' (Varela 1984b). It is the recognition of the conversation through different planes of interactions that makes Varela's theory a unique approach to cognition. He recognised that, if integrated in the biomechanical chemistry of the organism and, ultimately, if positioned within a wider environment, cognition is limited and, simultaneously, empowered in the morphodynamical constraints of their body. These cycles portrayed the organism as an entity in continual interplay that included not only cues coming from the environment, but also endogenous actions that prompted physiological responses (Varela 1979). This is why, for Varela, we should understand cognition as resulting from an integration of elements that, until then, we may have thought to be disjointed.

Successful cognition is, for Varela, essential for the overall well-being of an organism, suggesting an intimate connection between cognitive enablement and health. For an organism to gain the ability to draw cognitive significance to their surroundings and acknowledge their world as cognitively accessible, the correct communication between the elements taking part in this process must be ensured (Varela 1999). With Varela, the philosophical discussion starts to focus not only on elements that grant organisms cognitive capabilities, but also the mechanisms allowing the integration of said elements. He referred to this kind of multi-level interactions as 'coupling' (Varela 1981a) and suggested that failure to properly 'couple' could be detrimental for an organism. In consequence, he presents cognition as shaped through interactions from varied sources that coordinate and produce a meaningful order upon which organisms rely to achieve a general sense of well-being (Varela 1984a).

As such, Varela's efforts to underline the elements involved in cognition invite cognitive sciences to analyse the effect perturbations in any of the elements included in that multi-level network of interactions have in the coupling through which the organism accesses reality and, thus, the way in which the organism understands their world could be compromised if access to those elements were hindered, obscured, or prevented (Varela 1981b). It is based on this intuition that I will explore in Sects. 3 and 4 of this paper whether the menstrual cycle can be understood as an element in the process of cognitive enablement and if, consequently, the lack of understanding about the menstrual cycle could present a greater detriment for menstruating organisms that expected.

2.2 The Physiological, Psychosocial, and Environmental Realms of Cognition

With Varela offering an innovative take on cognition, the purpose of this paper, yet, requires a deeper understanding on the connection established between the elements

that prompt cognitive enablement. Varela recognised the multi-leveled nature of the process through which organisms access cognition but did not define how those levels articulate. Following Varela, Fuchs introduced mechanisms of coupling into his definition of lived experience, referring to cognitive enablement as a process of 'synchronisation' that, if successful, should generate a feeling of well-being in the organism (Fuchs and Schlimme 2009). Taking upon the task of establishing the levels that interact when attaining this synchronisation, Fuchs split the elements that lead an organism to cognitively understand their environment into two levels of interaction; a physiological level and a psychosocial level (Fuchs 2001).

On a physiological level, Fuchs observed organisms regularly attuned to bodily rhythms, among which he highlighted the circadian rhythm. These rhythms had the characteristic of being coordinated through timekeepers, endogenous or exogenous, that gave organisms the chance to, at any point of the cycle, attune with their pace. For Fuchs, these rhythms kept organisms' inner order from 'asynchrony', from falling into processes of 'anorganic nature' that could alter the correct development of fundamental bodily functions and trigger a discomforting experience within the organism. Despite being referred to as 'physiological', Fuchs' definition of this level presented the organism beyond their bodily limits, as one with their environment. He interpreted the cues and responses derived from the environment in which the organism interacts as if they were part of their own biochemistry and, as a result, suggested the boundary between organisms and their environment to be dynamic, flexible, and strongly interactive (Fuchs 2017).

When looking at the psychosocial level of the lived experience of an organism, Fuchs furthered synchrony to the social environment of the organism. Through everyday interaction with others, organisms are subject to a constant process of cognitive attunement that, if successful, results in an experience resonance (Fuchs 2012). For Fuchs, organisms are recurrently exposed to this social attunement and are involved with its rhythm every time they interact with their day-to-day dynamics.

Fuchs, thus, elaborated on the articulation of Varela's coupling, providing a deeper understanding on the elements contributing to the synchrony crucial for cognitive enablement. While synchrony comes in each level, fulfillment can only be attained if synchrony is reciprocated between both levels. Taking this two-leveled scheme as a starting point, I will, in later sections, analyse how the rhythms involved in the menstrual cycle could fit either of these levels.

2.3 Acknowledging the Vulnerability of the Organism to Desynchronisation

It is worth noting, nevertheless, that the concept of synchronisation elaborated comes with an innovative characteristic, entailing some kind of bidirectionality. Whilst organisms depend on physiological, psychological, and environmental cues to attune to certain rhythms, organisms also needed to actively perform activities that

reinforced the coordination to those cues to avoid potential disruptions. The possibility of disruption is, in addition, another characteristic that makes the concept of synchronisation attractive for the purpose of this paper. Physiological and psychosocial levels undermine each other when the synchronisation is hindered. For Fuchs, disruptions in the synchrony of these levels should be expected throughout the lived experience of an organism and, to minimise their consequences, organisms ought to put in place specific coping mechanisms that would allow them to resynchronise with their physiological and social environments (Fuchs 2010b). As suggested in the previous section, when referring to the desynchronisation of the organism, even if not explicitly, Fuchs recognised different degrees of asynchrony, depending on whether the desynchronisation is partial to one of the levels (Fuchs 2005). It is, indeed, in the desynchronised intersection of both levels that Fuchs identified the deepest degree of desynchronisation (Fuchs 2010a).

Underlining the possibility of unsuccessfully articulating physiological and psychosocial elements allowed for the exploration of the severe consequences asynchrony could entail. Failure to accurately synchronise with their environment, either through physiological or social elements, could lead the organism to a total exclusion from the coherence and resonance of their reality (Fuchs 2013b), plunging them into the 'experience of an unpleasant insistence of the body' involving cognitive disturbances reflected on feelings of heaviness, exhaustion, and restriction (Fuchs 2001). This lack of synchrony could lead to disordered cognitive capacities common in diagnosis of melancholic depression and schizophrenia (Fuchs 2005).

As such, the definition of a multi-level network allowed philosophers to recognise the complexity of the interactions enabling cognition, recognising that both endogenous and exogenous elements could provoke the disruption of physiological and psychosocial rhythms. Understanding the way organisms resonate with their reality, split into levels, therefore, is key to analyse not only synchronisation, but also desynchronisation, both presented as gradual phenomena.

2.4 Cognitive Enablement Through Participatory Sense-Making

Closing this first section and after a thorough overview of the vocabulary I will be using in further sections of the paper, it is key to analyse how the synchronisation prompted through these levels could contribute to cognitive enablement. For the completion of this task, I find limitations in Varela's and Fuchs' definitions. While Varela and Fuchs clearly identified the consequences of desynchronisation within the realm of clinical practice and, particularly, psychiatry, they did not account for more modest and yet disruptive grades of failed attunement. This is why, for this last part of the section, I will lean mainly on De Jaegher's concept of participatory sense-making.

This idea of total loss of cognitive enablement is explored in De Jaegher and Di Paolo, who paid special attention to the situation in which an organism is excluded from participation and, thus, is prevented from participating in activities of sense-making. Sense-making refers to the process through which organisms are incorporated to a meaningful frame (De Jaegher and Di Paolo 2007). Organisms have an active role in the tailoring of this complex network and actively participate in the process of generating meaning. Organisms do not take part in this activity isolatedly, but through a process of 'incorporation' in which meaning is generated through and needs to be ratified by social recognition (De Jaegher et al. 2010). De Jaegher and Di Paolo underlined the fragility of this sense-making process, highlighting a myriad of factors that are likely to compromise the performance of said activity and that could lead to the organism failing to resonate with it. Organisms are, in that sense, sensitive to losing the sense of coordination with their network and, if not successfully resynchronised with their reality, could be kept from participating in the activity of sense-making, furthering the gap with their environment. The parallelisms between the concept of incorporation in De Jaegher and Di Paolo and that of synchrony in Varela and Fuchs point at the activity of sense-making as the last step of the process of cognitive enablement needed for this paper.

The concepts summarised in this section establish the base for the discussion on cognitive enablement and the inclusion of physiological, social, and environmental elements in the manifold of interactions that allow organisms to access cognitive meaningfulness. It is through this vocabulary that this paper will enquire about the introduction of menstrual cycles among the rhythms involved in the attunement of menstruating organisms and the role it might have in the occurrence of cognitive disturbances.

3 The Relevance of Synchronisation in Biological Rhythm

According to this framework, organismal cognition requires synchronisation, to both physiological or psychosocial rhythms. This idea of synchronised rhythmicity is, however, not exclusive of philosophy, and neither is the intuition that synchronising to those rhythms would contribute to an overall well-being. It is within the concept of 'entrainment' in biology that philosophy finds clear similarities with the coupling that occurs between Varela's cycles of operations and Fuchs's levels. In this section, I will underline the complementarity of the narratives postulated in philosophy and biology regarding synchrony and entrainment and I will provide a picture of the biology behind menstrual rhythmicity.

3.1 The Evolutionary Predisposition to Synchrony

In biology, the process bodies undergo when synching internal functions to exogenous hints is referred to as entrainment (Kriegsfeld et al. 2002). Entrainment denotes a series of interactions through which independent processes become aligned and coordinate to adhere to a particular timing. The coupling cue bodies are most familiar with is daylight and, as a result, the rhythm based on which physiological functions have been most often analysed is the 24-hour light:dark cycle or, more accurately, the circadian rhythm (Satinoff 2001). This alignment, although reflected in a wide variety of bodily processes, is orchestrated by the hypothalamus, the endocrine organ, hosted in the brain, in charge of the regulation of hormone production and secretion. Due to its involvement with bodily processes, the knowledge on human biology is solidly based on the circadian rhythm and its timing.

From an evolutionary point of view, successfully adhering to environmental cues has been thought to be essential for the organism's well-being (Gerhart-Hines and Lazar 2015). The deterioration of the sleep-wake cycle, for instance, has been found to be a contributing factor to cognitive disorders such as generalised anxiety and depressive thoughts (Simon 2009; Gerhart-Hines and Lazar 2015). These disorders occur as a consequence of an interrupted entrainment, of a disrupted interaction between the human body and its cues and, ultimately, between the circadian rhythm and the hormones the rhythm controls (Beersma 2007). Dysfunctions caused by disturbances of the circadian rhythm are, ultimately, hormonal, and are linked to disruptions of the secretion pattern of different hormones, such as cortisol (Mohd Azmi et al. 2021) or melatonin (Reiter et al. 2020) often due to the interrupted, uncoordinated, or inhibited signaling of the hypothalamus. Far from having an impact on the secretion of a single hormone, the disruption of endocrine processes can precipitate a domino effect in other hormonal pathways.

Due to millenia of adaptation, organisms feel most comfortable when adhering to specific timings, ratified by periodical cues that are readily available in our environment, and that prompt them to follow a timing that is beneficial for them (Hut and Beersma 2011). Almost naturally, bodies would interact with those cues and adapt, accommodating a wide range of physiological processes to that schedule. For that synchrony to succeed, in contrast, human bodies need to actively adhere to practices compliant with such cues to avoid any disruptions. With the purpose of deepening into the research of these processes, disciplines presented under the umbrella term of 'chronobiology' have furthered the study of how bodies may support bodily rhythms through elements such as exercise or food (Postolache and Raheja 2016). Chronobiology has, however, considered the concepts of bodily rhythms and circadian rhythm equivalent, focusing most of its effort on a wide but yet not exhaustive range of rhythms.

3.2 A Second Biological Clock, the Infradian Rhythm

Against popular beliefs, the circadian rhythm does not account for the only rhythm all human bodies entrain with or, more accurately, it is not the only rhythm menstruating organisms entrain with (Koninck 1991; Stiller and Postolache 2005). Lack of focus on these non-circadian rhythms, nevertheless, led the scientific literature to extrapolate to menstruating organisms findings that did not account for a crucial part of their physiology (Simon 2005). It is to be asked if the lack of focus on menstruating organisms on chronobiology could have obscured the knowledge menstruating organisms have over the way their bodies synchronise with their, as well as whether menstruating organisms present any inherent rhythms derived from the secretion patterns of their sexual hormones.

In fact, the meaning of the word 'circadian' is not tied to that particular cycle, nor does it refer to a rhythm that is dominant or primary. The adjective 'circadian' is used to speak about any rhythms completed periodically within the span of a day and are present not only in animals, but also in plants, fungi, and bacteria (Dunlap and Loros 2017). Not all bodily processes are framed into a 24-hour deadline, even if they might have certain interactions with circadian cues. Recurrent cycles that are completed several times throughout the day and that, consequently, last less than 24 h are 'ultradian' rhythms; any cycles lasting over a day are, in contrast, referred to as 'infradian' rhythms (Pfaff 2018).

Human bodies present cycles of the three kinds and, while organisms' circadian nature might have been overly highlighted, the number of bodily processes adhering to ultradian and infradian rhythms is not to be ignored. Sexual hormones in menstruating organisms, for instance, are secreted in a more elongated period that, while popularly thought to be 28 days, can have any length between 21 and 35 days (Mihm et al. 2011; Bull et al. 2019). Thus, menstruating organisms, unlike non-menstruating organisms, involves a hormonal secretion pattern of sexual hormones that diverts from circadian timings and is aligned to infradian cues (Draper 2018).

The hormonal activity developed throughout the menstrual cycle is periodical and recurrent, but not consistent throughout the 21–35 days, and, as a result, can be differentiated in four phases or events based on the fluctuations in the concentrations of the sexual hormones that take part in the cycle, such as oestrogen, progesterone, the gonadotropin-releasing hormone (GnRH), the follicle stimulating hormone (FSH), and the luteinizing hormone (LH) (Hawkins and Matzuk 2008). None of the hormones mentioned are present in homogeneous levels along the menstrual cycle; lack of homogeneity should not, in contrast, be mistaken for lack of periodicity.

Menstruating organisms rhythmically advance through four physiologically distinct phases or events that are consecutively completed and that closely interact to fulfil their ultimate purpose, enabling reproduction. In their uniqueness, each phase has its own length and affects differently not only other physiological functions of the female body, but also its cognitive and emotional responses. Through medical

convention, it has been agreed that the first full day of bleeding is to be considered the start of the menstrual cycle and, consequently, menstruation is understood to be the first event or phase of the cycle. Cease of menses is followed by the second phase of the menstrual cycle, the follicular phase, that stimulates the growth of eggs or follicles in the ovaries. The maturation of the follicle prompts the event of ovulation, the phase through which the follicle will be fertile and can be fertilised. If not fertilised, the follicle deteriorates into the *corpus luteum*, a temporary endocrine body that will induce the thickening of the uterine lining. This lining will be shed in the next menstrual event, setting off the start of a new cycle. The length of each phase is thought to be predictable and consistent from cycle to cycle, with the menstrual event lasting from 2–7 days, the ovulatory event taking just over 24 h, and the luteal phase being 12–14 h. The length of the follicular phase is considered the most variable and is more likely to vary due to endogenous and exogenous causes, such as emotional, physiological, or environmental stress (Hawkins and Matzuk 2008; Barbieri 2014; Draper 2018).

As proof of the delicacy of the interactions between the hormones that coordinate the menstrual cycle, it should be underlined that the entirety of the menstrual cycle is conducted not only by the secretion of a single hormone, the gonadotropin-releasing hormone (GnRH), but through its secretion pattern (Thompson and Kaiser 2014). This hormone, produced in the hypothalamus, adapts the length and cadence of its pulses, signaling to the anterior pituitary if higher FSH or LH levels need to be released. Through low-frequency pulses, FSH is secreted, triggering the maturation of a follicle; through high-frequency pulses, LH is secreted to prompt the release of the matured follicle and, consequently, the event of ovulation (Marshall et al. 1991; Marshall et al. 1993). The disruption on the secretion frequency of the GnRH is common in modern society and could explain the increase in the incidence of cognitive disorders linked to the menstrual cycle, such as the PMDD (Premenstrual Dysphoric Disorder) (Tsutsumi and Webster 2009).

As a result, despite the lack of focus of research in the area of chronobiology on the rhythm involved in the menstrual cycle, this section closes with the argument that the menstrual rhythm is a rhythm crucial for the health of menstruating organisms, coordinated and supported through a careful and sophisticated myriad of physiological processes that, as I will explore in Sect. 3, also prompt psychosocial fluctuations that affect the daily life of menstruating organisms.

4 Benefits of Adapting Everyday Habits to Bodily-Rhythms

In Sect. 2, I elaborate on how menstruating organisms present a rhythm that has not been considered, neither in philosophy, nor within the research of chronobiology, and that, consequently, has not been incorporated to the understanding of bodily synchronisation postulated by philosophical work on embodied cognition. That section, ratifying Varela's and Fuchs' intuitions on the severe consequences asynchrony could have in organisms, also highlighted the physiological effects caused if

failing to synchronise with specific rhythmic cues. Underlining the relevance given to body-rhythms in the cognitive enablement of the organism, described in Sect. 1, it becomes urgent to question to what extent the lack of focus on the menstrual rhythm might have hindered the understanding available on menstruating organisms and put their well-being in jeopardy.

To account for the consequences that failing to understand menstrual rhythms as an essential rhythm might have had, I will underline the bidirectionality implied in the process of cognitive enablement overviewed in Sect. 1 and I will analyse of how rhythms affect organisms, not only physiologically, but also psychosocially, and how organisms ought to accommodate to those effects in pursuance of a higher degree of synchrony. This section will first account for the physiological and, mainly, psychosocial changes entailed to the circadian rhythm, moving onto those changes caused by the infradian rhythm of the menstrual cycle halfway through the section. Aware of the limited research conducted on the physiological and psychosocial effects mensal rhythms have on menstruating organisms, I will close the section with a case study in which adapting routines to the rhythmic fluctuations of the menstrual cycle proved to be beneficial.

4.1 Physiological and Psychosocial Changes Throughout Circadian Cycles

Circadian fluctuations are ingrained in the manner societies understand their lifestyle. Due to the daily secretion of hormones including cortisol and melatonin, the perception an organism has of their energy and vitality fluctuates in a 24-hour manner, peaking early in the morning and diminishing throughout the day, reaching its lowest point in the evening. The effect of these changes on the physiological and psychosocial capacities of organisms led social institutions to tailor their lifestyle to adopt schedules that had the purpose of optimising moments of maximum energy.

Prior to findings on chronobiology, popular beliefs were based on the intuition that circadian cues were connected to the changes in energy experienced throughout the day. Furthermore, peak energy levels were thought to prompt periods of higher cognitive performance, enhancing cognitive functions, such as perceptual cognitive speed, cognitive throughput, attention, and memory (Goel et al. 2013). This intuition successfully moulded the way in which everyday tasks were scheduled in modern and, most accurately, Western societies, particularly when looking at the scheduling of the working day, in which energy peaks fall within working hours. The performance improvement observed when organisms respected their circadian cadence and focused their peak-energy moments on their workday invited communities to understand all bodies as circadian bodies that would present energy variations that would adhere to a 24-hour span (Boubekri et al. 2020).

This process of circadian adaptation, developed a shared lifestyle that allows organisms to account for their circadian changes and adhere to a life-cadence that

respects them, facilitating the synchronisation of physiological and psychosocial processes. In addition, this lifestyle contributed to the avoidance of circadian disruptions that, as mentioned in Sect. 2, can cause generalised anxiety and depressive thoughts.

4.2 Cognitive Changes Throughout the Menstrual Cycle

Circadian adaptation is beneficial for menstruating organisms, certainly. However, while menstruating organisms follow circadian rhythms, they are also infradian bodies, with a series of fluctuations that take place in a period wider than a 24-hour day. These fluctuations impact on psychosocial functions and, due to the complexity of their interactions, result in a more intricate pattern that does not necessarily adhere to 24-hour peak-dip norm. In recent research, physiological (Farage et al. 2009), and psychosocial functions (Sundström Poromaa and Gingnell 2014) have been noted to change subject to the phase of the menstrual cycle menstruating organisms might be going through. Menstruating organisms would, indeed, experience circadian variations caused by the secretion of hormones whose production pattern corresponds to the 24-hour cycle, including cortisol, and melatonin, but the production and reception of those hormones would be embedded in a much more complex infradian framework and, thus, circadian fluctuations would be emphasised or, contrarily, inhibited.

If referred to, the changes linked to menstrual cycles have been categorised as not only unpredictable, but also erratic and irrational, discouraging any research to elaborate on them. Opposed to these presumptions, the menstrual cycle alters the manner in which menstruating organisms approach, experience, and interact with their own organism and the environment. Lack of research on infradian rhythms and, particularly, menstrual cycles, however, has caused popular societies to disregard the psychosocial changes prompted by this rhythm, leading to a situation where a rhythm that is embodied by over half of the population is absent from the current knowledge available on cognitive enablement.

As a result, literature on the psychosocial effects of the menstrual cycle is limited and has not been successfully incorporated into the glossary of common practice, preventing social routines from being adapted to any rhythms other than the circadian. To reduce this knowledge gap on how menstrual cycles might influence the cognitive enablement of menstruating organisms, I summarise in this section, sorted by phase, some of the most common physiological and, mainly, psychosocial fluctuations caused by the menstrual cycle.

During the menstrual phase, due to the reduction of the concentration of any kind of sexual hormones, menstruating organisms become susceptible to pain (Iacovides et al. 2015). As bleeding ceases, the cycle is restarted through the maturation of another follicle in the ovaries, triggered by the timely secretion of the FSH that, accordingly, also prompts a raise in the levels of available oestrogen (Barbieri 2014).

The surge of oestrogen characteristic of the follicular phase is known to boost mood and energy levels, with menstruating organisms feeling active and energised as they approach ovulation. Oestrogen has also been attributed psychological benefits, including enhanced memory, verbal skills, and a higher predisposition towards social and leisure activities (Sundström Poromaa and Gingnell 2014). With oestrogen promoting the secretion of stress-hormones (Babb et al. 2013), such as cortisol and adrenaline, as well as mood modulators such as endorphins, menstruating organisms perceive their bodies are better coordinated, appear to have faster reflexes, and feel, overall, more confident (Sundström Poromaa and Gingnell 2014). The stimulating effect oestrogen has over these stress hormones can, nevertheless, trigger feelings of restlessness, tension, or even anxiety in menstruating organisms more sensitive to their effect or that are oestrogen dominant and, as a result, present higher levels of oestrogen in their bodies (Harvey et al. 2009). It is within the last few days before ovulation that menstruating organisms experience a peak of testosterone (Barbieri 2014), increasing libido, competitiveness, and impulsivity. A decreasing trend in energy levels is observed mid-cycle, on average, around the day 14 of the menstrual cycle, after ovulation.

A significant dip in oestrogen levels after the ovulatory event is behind the changes menstruating organisms feel during the second half or their cycle, when the boosting effect of this hormone starts to fade and feelings of fatigue, lack of motivation, irritation, and overall sadness begin to increase (Li et al. 2020). If the follicle matured and released by the ovary is not fertilised, the egg is degraded into an endocrine structure referred to as the *corpus luteum* and prompts the last phase of the menstrual cycle before menstruation, the luteal phase. The luteal phase is identified through a surge of progesterone. Menstruating organisms going through this phase will experience their energy levels lowering as progesterone concentrations rise, due to the sedating nature of this hormone. Intense sensations of fatigue and tiredness can also be attributed to progesterone because of its impact in temperature, basal metabolic rate (BMR), and overall calorie expenditure of female bodies. The changes in metabolic responses prompted by progesterone also include enhanced hunger cues, a higher sensitivity to insulin, and drops in blood sugar, as well as temporary water retention (Ziomkiewicz et al. 2012), which can compromise body-confidence and self-esteem.

As menstruating organisms get closer to the end of their cycles, with progesterone having reached its peak and steadily decreasing, they report feelings of sadness and worry. Changes in the predisposition of menstruating organisms to take risks, try out new experiences, and interact in social events have also been observed in this phase, with menstruating organisms choosing to take part in situations they consider familiar, comfortable, and safe (Bröder and Hohmann 2003). This instinct to opt for more conservative choices has been thought to be linked to the fact that the immune response becomes impaired in menstruating organisms during the luteal phase and, as a consequence, menstruating organisms are at higher risk of developing infections and being attacked by viruses in this phase (Weinberg et al. 2011). The effects of this immunosuppression are not experienced until after the secretion of progesterone ceases and, subsequently, until the beginning of menstruation.

Consequently, menstruating organisms experience cyclical changes that correspond to a sophisticated rhythmicity, affecting their cognitive enablement. With lengthy research supporting the perks of furthering scientific knowledge on bodily rhythms and adapting everyday schedules to adapt to biological rhythms, it remains to be answered whether menstruating organisms might have been prevented from accessing the knowledge that could have led them to optimising their bodies' functioning, as well as if the lack of understanding on their physiology might have disrupted, impaired, or even chronically threatened their wellbeing. Moreover, in the same way failing to follow lifestyles that match the cadence of circadian rhythms could prompt diagnoses of depression and anxiety, the limited adaptation of the routines of menstruating organisms to the menstrual rhythm urges the question whether these bodies are put in a higher risk of developing these pathologies.

4.3 The Benefits of Adopting Infradian Routines

Conscious of the limited nature of research on the psychosocial fluctuations of menstruating organisms, recent lines of study have chosen to look at the effects adapting some of the daily activities of menstruating organisms to their infradian changes can have on the perception of their well-being. These studies have primarily focused on the changes menstruating organisms perceive when looking at their sports performance throughout their menstrual cycle (McNulty et al. 2020).

Research on the infradian adaptation of the activity of menstruating organisms represents the first attempt of scientific literature to understand the benefits that could come from synchronising menstruating organisms to one of their primary body-rhythms. This line of research, focused on athletic performance, suggested breaking away from week-based progressive training schedules, prompting a meaningful shift in the way current sports science understands body composition, muscle-growth, and overall health. This research also endorsed claims denouncing the exclusion of menstruating organisms from sport sciences, pointing out the fact that the majority of studies in exercise and performance had been done focusing exclusively on non-menstruating organisms and, indeed, male physiology, extrapolating research findings to menstruating organisms, even if such results had not been ratified. The fluctuations lengthy covered in previous paragraphs of Sect. 3, in fact, reveal that, far from extrapolable, menstruating organisms present not only physiological, but psychosocial characteristics that impact their day-to-day and that, for menstruating athletes, could be key for optimal performance.

According to that body of research, the main phases of the menstrual cycle, the follicular phase and the luteal phase, involved exceptional metabolic changes (Draper 2018). During the follicular phase, oestrogen creates a metabolic situation ideal for muscle-growth and repair, lowering the BMR, enabling levels of cortisol and adrenaline to raise without disrupting general stress-levels, and increasing the energy of menstruating organisms (Bisdee et al. 1989). In this phase, female bodies

experience an improved perception of their performance in HIIT or higher intensity exercise.

As ovulation happens, the testosterone surge improves performance in endurance sports, with athletes feeling more comfortable practising exercises that require them to sustain their pace for prolonged periods of time.

Through the luteal phase, the metabolism of menstruating organisms speeds up, together with their BMR, and the nutritional needs increase in 100 to 300 calories regardless of physical expenditure (Bisdee 1989). This increase in metabolic speed can prompt menstruating organisms to feel more tired than usual. Progesterone, the main hormone of the luteal phase, increases the pain threshold of menstruating organisms and mitigates fatigue cues, helping athletes reach better marks in anaerobic sports that involve resistance (Ziomkiewicz et al. 2012).

Research conducted on the impact of the menstrual cycle in sports performance perception, while acknowledging the differences that are observed from menstruating subject to menstruating subject, pleads for a training environment that considers the cyclical nature of menstruating organisms (McNulty et al. 2020). Adapting coaching schedules and exercise intensity to their menstrual cycle was proven to not only improve the perception of the performance of menstruating organisms, but also helped them gain a better understanding of the physiological and psychosocial changes that they undergo throughout their menstrual cycle. Efforts in this field of research have encouraged sport professionals to openly include athletes' menstrual cycles in training conversations, modifying the cadence of their exercise calendars to accommodate their menstrual rhythm (Julian and Sargent 2020).

5 The Desynchronised Menstruating Organism

In Sect. 3, I have provided an in-depth account of the cognitive changes menstruating organisms perceive as they advance through the menstrual cycles. Similar to the psychosocial fluctuations connected to circadian rhythms overviewed at the beginning of Sect. 3, menstruating organisms experience variations in the perception of their energy level in an infradian fashion. This previous section exemplified how adapting daily activities to the circadian rhythm had been observed to improve performance perception and overall sense of well-being, suggesting that mimicking that practice and matching the schedules of menstruating organisms to their menstrual rhythm could be beneficial for their well-being.

If listening to bodily rhythms, be them circadian or infradian, contributes to the general well-being of an organism, the question of how excluding one of those rhythms from popular and institutional considerations could have impacted the cognitive enablement of menstruating organisms seems unavoidable. Addressing these findings through the terminology of the embodied cognition postulated in Sect. 1, I will argue in this section that menstruating organisms could be more vulnerable to hindered synchronisation.

5.1 The Consequences of a Desynchronised Menstruating Organism

As explored in Sect. 2, having a lifestyle that systematically antagonises circadian cues is known to have critical consequences in an organism's physiology, impacting mechanisms that include but are not limited to those connected to sleep. Circadian disruptions can result in physiological and psychosocial dysregulations, compromising learning, memory (Gibson et al. 2010), and affective regulation (Samuels and Hen 2011).

Studies in sport performance analysed at the end of Sect. 3 identified various consequences ignoring the effect the menstrual cycle had in menstruating athletes, including increased feelings of tiredness or anxiety. Recent studies have pointed at this antagonisation menstruating organisms are exposed to when they are prevented from following a lifestyle that could better correspond to their body-rhythms as a potential reason behind the rapidly increasing incidence of menstruating organisms with hormonal disturbances. When compared to the number of non-menstruating organisms that suffer from some kind of endocrine imbalance, the proportion of menstruating organisms whose bodies find it challenging to keep up with a healthy hormonal profile is remarkably high (Golden et al. 2009; Lauretta et al. 2018; Crafa et al. 2021).

Among the menstrual disorders most commonly reported, the disorder that would more closely resemble the examples of pathologies that Varela postulated when looking at diagnoses resulting from unsuccessful synchronisation would be the premenstrual dysphoric disorder (PMDD). PMDD is a broadly defined pathology that entails a number of psychosocial symptoms that could be considered disrupting, such as depressed mood, anxiety, and irritability, sometimes even leading to episodes of paranoia (Hantsoo and Epperson 2015). Despite the lack of presence of this disorder in popular conversations and common knowledge, it is thought that, on average, 5–8% of menstruating organisms suffering from severe premenstrual symptoms could fulfill the criteria of being diagnosed with PMDD (Yonkers et al. 2008).

Looking at these disturbances through the tools explored in Sect. 1, the psychosocial disruptions observed could be interpreted as resulting from a failed synchronisation.

5.2 Stigmatising Menstruating Organisms and Effects on the Promotion of Menstrual Awareness

As explored in Sect. 1, desynchronisation is an inherent part of the cognitive enablement of organisms. It is not plausible to think of a lived experience that does not present any degree of desynchronisation. In fact, bodies are subject to periodical desynchronisation that deteriorates their attunement with body-rhythms,

in both a physiological and psychological level (Fuchs 2013a). These periodical instances of desynchronisation are connected to a lack of correspondence with particular cues and, should bodies insist on following those disruptive cues, they would be in an increased danger of falling into a state of asynchrony (Fuchs 2013b). As such, without a wide understanding of their menstrual rhythmicity, the possibility of menstruating organisms to be in synchrony with their bodies seems to be at a strong disadvantage.

The infradian rhythm derived from the menstrual cycle is not only disregarded by health professionals, but also avoided in popular and informal conversations. Despite having, even if limited, available research that proves that menstruating organisms experience noteworthy variations that can be explained through determined physiological and psychosocial variations, these bodies keep being addressed by institutions and society as unpredictable, unreliable, and erratic (Gottlieb 2020; Johnston-Robledo and Chrisler 2020). Menstruating organisms are brought up to perceive their bodies as unexplainable and, since they are not facilitated the access to the information regarding the perfectly coherent and steady changes that they go through during their menstrual cycle, they end up feeling alien of their own bodies, uncomfortable with their bodily reality, and unattuned to a rhythm that orchestrates an essential part of their physiology.

In fact, more often than not, menstruating organisms are addressed as if they were 'out of control'. Because of the limited access menstruating organisms have to general information about their physiology that could help them understand the cognitive changes they periodically go through during their menstrual cycle (Sundström Poromaa and Gingnell 2014), menstruating organisms themselves find the variations they perceive random, lacking coherence, and, ultimately, 'disordered' when compared to the circadian rhythm all bodies are supposed to be able to follow comfortably. The feeling of frustration is even stronger when menstruating organisms schedule their lifestyle to be strictly compliant with circadian cues and, yet, keep experiencing their bodies to vary in ways that not only are not recognised by those cues, but that are also not acknowledged within their social environment.

As seen in Sect. 3, current efforts to widen the research on the cognitive enablement of menstruating organisms, while receiving more attention than in the past, have been planned within frameworks that keep reproducing biased practices. The majority of recent studies on the psychosocial changes involved in the menstrual cycle have tended to be focused on how these variations might impact reproduction and, particularly, sexual arousal. Reducing the effect of menstrual fluctuations to merely reproduction has been a recurrent mistake in scientific research, failing to acknowledge the effect these hormones can have beyond the gonads. Likewise, most studies have tailored the requirements of their participants to filter out menstruating organisms through psychiatric interviews, ensuring that studies are solely performed on 'healthy' menstruating organisms. Consequently, menstruating organisms experiencing some kind of psychological disturbance are excluded from these studies (De Bruin 1999; Simon 2005; Holdcroft 2007; Mazure and Jones 2015; Liu and Di Prietro 2016), obscuring the crucial role the menstrual fluctuations could have in these disorders.

As a result, future research on the cognitive capabilities of menstruating organisms should consider the variations prompted by menstrual fluctuations as a relevant factor for their overall well-being. Similarly, research should include menstruating organisms with psychosocial disorders, furthering the knowledge on the causal role menstrual cycles could have in them. Due to the lack of precedent of research on menstrual rhythms, any efforts to establish new research should first review the protocols that have, until now, been followed, enquiring whether those protocols could be, in any way, biased.

5.3 Resynchronising Through the Menstrual Participatory Sense-Making

Reclaiming menstrual awareness should not, however, be understood as an effort limited to research bodies and academia. The knowledge over the menstrual cycle and the fluctuations linked to it remain essential for the synchronisation of menstruating organisms with their bodies and, consequently, for their cognitive enablement. To close this section, I will elaborate on the pathways available to menstruating organisms for the resynchronisation with their body-rhythm furthering on the terminology postulated by Fuchs, De Jaegher, and Di Paolo on participatory sense-making and, particularly, on the process of mutual incorporation necessary for accessing activities of participatory sense-making.

For De Jaegher and Di Paolo, the process of mutual incorporation is an expansion of the lived experience of each of the participants that corresponds to the reciprocal coordination of the factors that enable the embodiment of the participants with their bodies (De Jaegher and Di Paolo 2007) and, as a result, allow them to access activities of participatory sense-making. For De Jaegher, mutual incorporation does not only facilitate the establishment of a joint understanding of a shared reality, but it also introduces the subjects involved in this procedure into new domains of sensibility (De Jaegher et al. 2016).

Fuchs, De Jaegher, and Di Paolo defined this process to require not only for the subject to understand the rhythms they aimed to resynchronise with, but also for its social network to have a certain degree of awareness of such rhythms (De Jaegher and Di Paolo 2008a; Fuchs and De Jaegher 2009). For the success of the reincorporation of the subject to be guaranteed, the process must be shared with and ratified by, at least, another subject. Rhythms that are not widely shared are, consequently, considered harder to regain attunement with and, thus, bodies with these body-rhythms could find it particularly difficult, lonely and even self-questioning to regain attunement to a body-rhythm that is denied, contradicted, and invalidated.

The mutual incorporation postulated contributes to the cognitive enablement of bodies in several ways since it does not only allow subjects to create a shared awareness of their body-rhythms, but it also prompts them to acknowledge

divergences in their now joint understanding of their reality, forcing them to reconsider any mismatches found in this process. De Jaegher and Fuchs do not elaborate on the situation where mismatches are identified. Yet, it seems likely that socially shared assumptions could compromise this process (De Jaegher et al. 2010). In all likelihood, if two subjects were to contribute to a process of mutual incorporation with opposite or divergent understandings of reality, the account with wider social endorsement would be prioritised over the other, leaving little to no room to the inquiry of dominant narratives (De Jaegher and Froese 2009). Hence, non-dominant or recurrently marginalised narratives, such as menstrual narratives, could encounter difficulties to not only enter, but succeed at making an impact in conversations of mutual incorporation.

For situations in which this mutual incorporation is not available or, as referred above, in which the procedure of mutual incorporation fails to acknowledge less widely endorsed accounts, De Jaegher and Fuchs postulated the possibility of performing a sort of unidirectional incorporation, although they did not recommend it (Fuchs and De Jaegher 2009). This procedure focuses on the familiarisation of a subject with a particular rhythm through activities that involve training or learning exercises, as well as research. Attuning to their bodily rhythms individually would, however, always remain exposed to potential disruptions, since the synchrony that the body perceives is based upon their coordination with rhythms that are not socially shared and that might be consistently put into question.

Due to its inherently intimate nature, the menstrual cycle could be a perfect phenomenon to coordinate *to*, to acknowledge through unidirectional incorporation. Even if not recommended as the only type of incorporation through which bodies might interact with reality, De Jaegher and Fuchs recognised this kind of incorporation to be a step towards the correct direction (Fuchs and De Jaegher 2009). Approaching menstrual cycles through practices including research, active learning, and training, even in unidirectional, would represent an instance of incorporation that allows the body to transcend itself and, even if partly, to merge with the environment.

The difference between these processes of incorporation should, however, remain clear. While the subject coordinates *to* when engaged in a unidirectional incorporation, mutual incorporation involves coordination *with.*

When comparing both types of incorporation, De Jaegher and Fuchs underlined in mutual incorporation a degree of autonomy and otherness that could not be attained through unidirectional coordinations. Similarly, the process of identifying mismatches would only truly be accessible when in interaction with others. Following these differences drawn by De Jaegher and Fuchs, it could be theorised that, bringing Fuchs' levels into the equation that, for bodies to feel in synch and become cognitively enabled, it would be necessary not only to train, learn, or research for their physiological attunement to be successful, but it would also require that attunement to be shared and, at least, partially endorsed by their social environment in with which bodies will coordinate (Fuchs and De Jaegher 2009). This need for a shared awareness becomes even more crucial when looking at the lack of research on

menstruating organisms, limiting the access to the few clinical resources concerning menstruating health.

5.4 A New Domain of Menstrual Significance

For menstruating organisms to be cognitively enabled, consequently, an understanding of their menstrual cycle should be promoted not only from a perspective of self-knowledge and reflection, but from a social address. This should demand society and, most importantly, the health and educational institutions to have a proactive approach towards menstruating organism-rhythms. Menstruating organisms should be entitled access to scientific resources essential to solidify their awareness over their bodies, as well as to a community that would share that awareness and recognise the fluctuations resulting from these body-rhythms as just as crucial as the fluctuations brought by circadian cues.

The purpose of extending the understanding of menstruating physiology to society should not be that of validating any experience linked to the menstrual cycle, but to prompt a shared understanding in which subjects coordinate with each other, putting into the spotlight mismatches to be studied and accordingly acknowledged. These interactions could culminate into a socially shared understanding of menstruating health that demands political, health, and educational institutions to include these body-rhythms in their protocols, plans, and conversations.

For De Jaegher and Fuchs, these interactions could generate meaning, that, even if not directly experienced by a particular subject, can provide insight into the embodiment of the other, improving the understanding of subjects that do not share those rhythms with menstruating organisms (Fuchs and Froese 2012). With processes that only involve menstruating organisms being politically recognised, with health care institutions considering body phenomena present in menstruating organisms as worth researching and, with educational institutions cultivating menstruating organisms from their childhood to understand their body-rhythm, a completely different sphere of sense-making opens for menstruating organisms.

Consequently, once menstruating organisms are aware of the cognitive changes that occur throughout their menstrual cycle, it is likely that part of the discomfort, unfamiliarity, and, ultimately, anxiety that they feel due to those variations could disappear or, at least, be mitigated. This innovative understanding of menstrual rhythmicity also invites future research to consider the beneficial effect establishing routines that adhere to menstruating organisms' rhythm could have in lowering the incidence of disorders connected to the menstrual cycle, such as the premenstrual syndrome and the PMDD.

6 Conclusion

This paper argued that the synchrony of menstruating organisms with their menstrual rhythm is key for their cognitive enablement.

In Sect. 1, I summarised the conceptual glossary Varela, Fuchs, De Jaegher and Di Paolo have produced in the field of cognition, essential for the recognition of the process of cognitive enablement as embedded in a multi-leveled manifold of interactions that organisms perform with themselves or with others. In this paradigm of cognition, the organism's well-being is subject to them successfully coupling with their body-rhythms, rhythms that, complementarily, are affected by physiological and psychosocial cues prompted by their physical and social environment. Hence, menstrual rhythms become essential for the well-being of menstruating organisms.

In Sect. 2 of this paper, I have explored the phenomenon of synchronisation from the perspective of biological sciences, where the concept of synchrony finds its biological equivalent: the phenomenon of entrainment. I underlined how the research on entrainment, mainly conducted in the area of chronobiology, has focused on circadian rhythms, neglecting entrainment prompted by other body-rhythms, such as that entailed to the menstrual cycle. While menstruating organisms follow circadian variations, modulated by the secretion of hormones such as cortisol, insulin, and melatonin, the metabolism of such hormones and, consequently, their physiological and psychosocial effect, depends on an infradian framework.

In Sect. 3, I furthered the discussion on the consequences executing actions in discordance with body-rhythms have for organisms, prompting feelings of discomfort and desynchronisation, triggering depressive and anxious feelings, and, ultimately, jeopardising the access organisms have to their cognition. This section first focused on the negative effects of circadian disturbances, highlighting how societies and, particularly, Western societies, have developed through recent history institutionalised and widely shared routines that contribute to the following an everyday rhythm compliant to circadian cues. Analysing the limited literature on the physiological but, mainly, psychosocial fluctuations prompted by the menstrual cycle, in Sect. 3 I have also provided a detailed account of the physiological and psychosocial effects of hormones involved in the menstrual cycle, including as oestrogen, progesterone, GnRH, FSH, and LH, effects that can be further classified in events or phases referred to as menstruation, the follicular phase, ovulation, and the luteal phase. Unlike with circadian rhythms, I denounced in the last part of Sect. 3 the lack of institutionally developed routines, designed to contribute the adherence of menstruating organisms to one of their primary rhythms. To prove the benefits coming from this adaptation, I presented the findings of recent research on training routines adapted to better match the menstrual rhythm of menstruating athletes.

Lastly, in Sect. 4, I articulated the disruptions lack of awareness over menstrual rhythmicity could have prompted for menstruating organisms through the terminology elaborated on Sect. 1, aiming to develop throughout the section pathways menstruating organisms could have available to mitigate the effects or altogether avoid the effects of desynchronisation with their body-rhythm. In this section, the

phenomenon of synchronisation is presented as only successfully exercised in coordination, either *to* or *with,* granting an organism meaningful access to the realm of participatory sense-making. Organisms that are either unaware of or kept from this sense-making exercise are, consequently, highly vulnerable to states of uneasiness, health disturbances, and social ostracism. This section brings into focus the lack of information available to menstruating organisms and how this might have prevented menstruating organisms from accessing an accurate account of their own bodies, affecting menstruating organisms' cognitive enablement.

As a result, this paper considers crucial for future research in cognition to develop a wider understanding of the infradian rhythm. This paper invites further research on the role of the menstrual cycle in the cognitive well-being of menstruating organisms and calls for initiatives for menstruating organisms to better connect with their menstrual rhythm that should, consequently, be institutionally endorsed in scientific, political, and social spheres. This paper represents one of the first but surely not the last step towards a body of research focused on the production of a shared understanding of the menstruating organism that involves a thorough research of their body-rhythms and results in the establishment of clinical protocols and social policies that enable menstrual awareness and empower menstruating organisms.

Acknowledgements This paper would not have been possible without the support of my supervisor, Arantza Etxebarria. Her guidance and attention to detail were key during the production of the final draft of this paper. My PhD peers at the University of the Basque Country were also kind enough to look over my drafts and helped me polish some last detail. I am also grateful for the insightful comments offered by the anonymous peer reviewers at Springer. Their advice and knowledge have improved this paper in innumerable ways and I am thankful for that.

References

Babb JA, Masini CV, Day HE, Campeau S (2013) Stressor-specific effects of sex on HPA axis hormones and activation of stress-related neurocircuitry. Stress 16(6):664–677

Barbieri RL (2014) The endocrinology of the menstrual cycle. Methods Mol Biol 1154:145–169

Beersma DGM (2007) Circadian control of the sleep-wake cycle. Physiol Behav 90(2–3):190–195

Bisdee JT (1989) Metabolic changes during the menstrual cycle. Br J Nutr 61:641–650

Bisdee JT et al (1989) Changes in energy expenditure during the menstrual cycle. Br J Nutr 61:187–199

Boubekri M et al (2020) The impact of optimized daylight and views on the sleep duration and cognitive performance of office workers. Int J Environ Res Public Health 17(9):3219

Bröder A, Hohmann N (2003) Variations in risk taking behavior over the menstrual cycle: an improved replication. Evol Hum Behav 24(6):391–398

Bull JR et al (2019) Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. NPJ Digital Medicine 2(1):1–8

Crafa A et al (2021) The burden of hormonal disorders: a worldwide overview with a particular look in Italy. Front Endocrinol 12:694325

De Bruin DA (1999) Justice and the inclusion of women in clinical studies: a conceptual framework. In: Women and health research: ethical and legal issues of including women in clinical studies, vol 2 (Workshop and commissioned papers). National Academies Press, Washington

De Jaegher H (2009) Social understanding through direct perception? Yes, by interacting. Conscious Cogn 18(2):535–542
De Jaegher H, Di Paolo E (2007) Participatory sense-making: an enactive approach to social cognition. Phenomenol Cogn Sci 6(4):485–507
De Jaegher H, Di Paolo E (2008a) Making sense in participation: an enactive approach to social cognition. Emerging Commun 10:33
De Jaegher H, Di Paolo E (2008b) Enacting intersubjectivity: a cognitive and social perspective on the study of interactions, vol 10. IOS Press, Amsterdam, pp 33–47
De Jaegher H, Di Paolo E, Gallagher S (2010) Can social interaction constitute social cognition. Trends Cogn Sci 14(10):441–447
De Jaegher H, Peräkylä A, Stevanovic M (2016) The co-creation of meaningful action: bridging enaction and interactional sociology. Philos Trans R Soc B: Biol Sci 371(1693):20150378
De Jaegher H, Froese T (2009) On the role of social interaction in individual agency. Adapt Behav 17(5):444–460
De Koninck J (1991) Les rythmes biologiques liés au sommeil et l'adaptation psychologique [Biological rhythms associated with sleep and psychological adjustment]. J Psychiatry Neurosci 16(3):115–122
Di Paolo E (2005) Autopoiesis, adaptivity, teleology, agency. Phenomenol Cogn Sci 4(4):429–452
Di Paolo E, Fuchs T (2015) Toward an embodied science of intersubjectivity: Widening the scope of social understanding research. Front Psychol 6:234
Draper CF et al (2018) Menstrual cycle rhythmicity: metabolic patterns in healthy women. Sci Rep 8(1):1–15
Dunlap JC, Loros JJ (2017) Making time: conservation of biological clocks from fungi to animals. Microbiol Spectr 5(3):5–3
Farage MA, Neill S, MacLean AB (2009) Physiological changes associated with the menstrual cycle: a review. Obstet Gynecol Surv 64(1):58–72
Fuchs T (2001) Melancholia as a desynchronisation: towards a psychopathology of interpersonal time. Psychopathology 34(4):179–186
Fuchs T (2005) Corporealized and disembodied minds: a phenomenological view of the body in melancholia and schizophrenia. Philos Psychiatry Psychol 12(2):95–107
Fuchs T (2010a) Temporality and psychopathology. Phenomenol Cogn Sci 12(1):75–104
Fuchs T (2010b) Phenomenology and psychopathology. In: Schmicking D, Shaun G (eds) Handbook of phenomenology and cognitive science. Springer, New York, pp 546–573
Fuchs T (2011) The brain: a mediating organ. J Conscious Stud 18(7–8):196–221
Fuchs T (2012) Embodied affectivity: on moving and being moved. Front Psychol 5:508
Fuchs T (2013a) The phenomenology and development of social perspectives. Phenomenol Cogn Sci 12(4):655–683
Fuchs T (2013b) Existential vulnerability: towards a psychopathology of limit situations. Psychopathology 46(5):301–308
Fuchs T (2017) Ecology of the brain: the phenomenology and biology of the embodied mind. Oxford University Press, Oxford
Fuchs T (2019) The interactive phenomenal field and the life space: a sketch of an ecological concept of psychotherapy. Psychopathology 52(2):67–74
Fuchs T, De Jaegher H (2009) Enactive intersubjectivity: participatory sense-making and mutual incorporation. Phenomenol Cogn Sci 8(4):465–486
Fuchs T, Froese T (2012) The extended body: a case study in neurophenomenology of social interaction. Phenomenol Cogn Sci 11(2):205–235
Fuchs T, Schlimme JE (2009) Embodiment and psychopathology: a phenomenological perspective. Curr Opin Psychiatry 22(6):570–575
Gerhart-Hines Z, Lazar MA (2015) Circadian metabolism in the light of evolution. Endocr Rev 36(3):289–304
Gibson EM et al (2010) Experimental 'jet lag' inhibits adult neurogenesis and produces long-term cognitive deficits in female hamsters. PLoS One 5(12):e15267

Goel N, Basner M, Rao H, Dinges DF (2013) Circadian rhythms, sleep deprivation, and human performance. Prog Mol Biol Transl Sci 119:155–190
Golden SH et al (2009) Clinical review: prevalence and incidence of endocrine and metabolic disorders in the United States: a comprehensive review. J Clin Endocrinol Metab 94(6): 1853–1878
Gottlieb A (2020) Menstrual taboos: moving beyond the curse. In: Bobel C, Winkler IT, Fahs B et al (eds) The Palgrave handbook of critical menstruation studies. Palgrave Macmillan, pp 143–115
Hackney AC, Kallman AL, Aggön E (2019) Female sex hormones and the recovery from exercise: menstrual cycle phase affects responses. Biomed Hum Kinet 11(1):87–89
Hantso L, Epperson CN (2015) Premenstrual dysphoric disorder: epidemiology and treatment. Curr Psychiatry Rep 17(11):1–9
Harvey AT, Hitchcock CL, Prior JC (2009) Ovulation disturbances and mood across the menstrual cycles of healthy women. J Psychosom Obstet Gynaecol 30(4):207–214
Hawkins SM, Matzuk MM (2008) The menstrual cycle: basic biology. Ann N Y Acad Sci 1135(1): 10–18
Holdcroft A (2007) Gender bias in research: how does it affect evidence based medicine? J R Soc Med 100(1):2–3
Hut RA, Beersma DGM (2011) Evolution of time-keeping mechanisms: early emergence and adaptation of photoperiod. Philos Trans R Soc Lond, B, Biol Sci. 366(1574):2141–2154
Iacovides S, Avidon I, Baker FC (2015) Does pain vary across the menstrual cycle? A review. Eur J Pain 19(10):1389–1405
Johnston-Robledo I, Chrisler JC (2020) The menstrual mark: menstruation as social stigma. In: Bobel C, Winkler IT, Fahs B et al (eds) The Palgrave handbook of critical menstruation studies. Palgrave Macmillan, pp 181–119
Julian R, Sargent D (2020) Periodisation: tailoring training based on the menstrual cycle may work in theory but can they be used in practice? Sci Med Footb 4(4):253–254
Kriegsfeld L, Silver R, Gore AC, Crews D (2002) Vasoactive intestinal polypeptide contacts on gonadotropin-releasing hormone neurons increase following puberty in female rats. J Neuroendocrinol 14(8):685–690
Lauretta R et al (2018) Gender in endocrine diseases: role of sex gonadal hormones. Int J Endocrinol
Li SH, Lloyd AR, Graham BM (2020) Physical and mental fatigue across the menstrual cycle in women with and without generalised anxiety disorder. Horm Behav 118:104667
Liu KA, Dipietro NA (2016) Women's involvement in clinical trials: historical perspective and future implications. Pharm Pract 14(1):708
Marshall JC et al (1991) Gonadotropin-releasing hormone pulses: regulators of gonadotropin synthesis and ovulatory cycles. Recent Prog Horm Res 47:155–187
Marshall JC et al (1993) GnRH pulses—the regulators of human reproduction. Trans Am Clin Climatol Assoc 104:31–46
Mazure CM, Jones DP (2015) Twenty years and still counting: including women as participants and studying sex and gender in biomedical research. BMC Womens Health 15(1):1–16
McNulty KL et al (2020) The effects of menstrual cycle phase on exercise performance in eumenorrheic women: a systematic review and meta-analysis. Sports Med 50(10):1813–1827
Mihm M, Gangooly S, Muttukrishna S (2011) The normal menstrual cycle in women. Anim Reprod Sci 124(3–4):229–236
Mohd Azmi NAS et al (2021) Cortisol on circadian rhythm and its effect on cardiovascular system. Int J Environ Res Public Health 18(2):67
Pfaff DW et al (2018) Hormonal secretions and responses are affected by biological clocks. In: Pfaff DW et al (eds) Principles of hormone/behavior relations. Academic Press, London, pp 293–314
Postolache TT, Raheja UK (2016) Body rhythms/biological clocks. In: Encyclopaedia of mental health, vol 1. Elsevier, London, pp 193–203

Reiter RJ, Rosales-Corral S, Sharma R (2020) Circadian disruption, melatonin rhythm perturbations and their contributions to chaotic physiology. Adv Med Sci 65(2):394–402
Satinoff E (2001) Circadian rhythms. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social and behavioural sciences. Pergamon, Amsterdam
Samuels BA, Hen R (2011) Neurogenesis and affective disorders. Eur J Neurosci 33(6):1152–1159
Simon EK (2009) Psychiatric disorders associated with disturbed sleep and circadian rhythms. In: Squire LR (ed) Encyclopedia of neuroscience. Academic Press, London, pp 1167–1185
Simon V (2005) Wanted: women in clinical trials. Science 308(5728):1517
Stiller JW, Postolache TT (2005) Sleep-wake and other biological rhythms: functional neuroanatomy. Clin Sports Med 24(2):205–235
Sundström Poromaa I, Gingnell M (2014) Menstrual cycle influence on cognitive function and emotion processing-from a reproductive perspective. Front Neurosci 124(8):380
Thompson IR, Kaiser UB (2014) GnRH pulse frequency-dependent differential regulation of LH and FSH gene expression. Mol Cell Endocrinol 385(1–2):28–35
Thompson E, Varela FJ (2001) Radical embodiment: neural dynamics and consciousness. Trends Cogn Sci 5(10):418–425
Tsutsumi R, Webster NJ (2009) GnRH pulsatility, the pituitary response and reproductive dysfunction. Endocr J 56(6):729–737
Varela FJ (1979) Principles of biological autonomy. Elsevier North-Holland, New York
Varela FJ (1981a) Autonomy and autopoiesis. In: Roth G, Schwegler H (eds) Self-organizing systems: an interdisciplinary approach. Campus Verlag, Frankfurt, pp 14–24
Varela FJ (1981b) Describing the logic of the living: the adequacy and limitations of the idea of autopoiesis. In: Zeleny M (ed) Autopoiesis: a theory of living organization. North-Holland, New York, pp 36–48
Varela FJ (1984a) Living ways of sense-making: a middle path for neuroscience. In: Varela F, Livingstone P (eds) Disorder and order: proceedings of the Stanford international symposium. Anma Libri, Saratoga, pp 208–224
Varela FJ (1984b) Two principles for self-organization. In: Ulrich H, Probst GJ (eds) Self-organization and management of social systems. Springer, Berlin, Heidelberg, pp 25–32
Varela FJ (1999) First-person methodologies: why, when and how. J Conscious Stud 6:2–3
Varela FJ, Thompson E, Rosch E (2017) The embodied mind. In: Cognitive science and human experience, Revised edn. MIT Press, Cambridge, MA
Yonkers KA, O'Brien S, Eriksson E (2008) Premenstrual syndrome. Lancet 371(9616):1200–1210
Weinberg A et al (2011) Effect of menstrual cycle variation in female sex hormones on cellular immunity and regulation. J Reprod Immunol 89(1):70–77
Ziomkiewicz A et al (2012) Higher luteal progesterone is associated with low levels of premenstrual aggressive behavior and fatigue. Biol Psychol 91(3):376–382

Part II
Evolution, Language and Culture

Is Cultural Selection Creative?

Malena León

People don't invent complex tools, populations do. (Boyd et al. 2013, p. 3).

Abstract This paper aims to draw some theoretical relationships between two fields of research that have remained more separated than they should have: theories of creativity and theories of cultural evolution. Particularly, it argues that the mechanisms of cultural selection postulated by cultural evolutionary theories can make a hitherto neglected contribution to explanations of human creativity. To that end, I extrapolate the arguments in favour of the creativity of natural selection and weigh its applicability in the field of culture.

Keywords Creativity · Cultural evolution · Creativity of natural selection · Cultural selection

1 Introduction

The intelligent design argument (Paley 1802; Sober 2019) used for proving the existence of God can be reconstructed as follows:

(i) Biological organisms are designed objects.

(ii) Every designed object has a designer.

Therefore,

(iii) Biological organisms have a designer.

Dennett (1995) has argued that Darwin's theory provides sufficient means to refute the second premise, given that natural selection and not God would be the process responsible for the design. A 'cultural version' of the argument would apply almost trivially to cultural items, with the necessary modifications in the first premise and the conclusion (exchanging 'biological organisms' for 'cultural items'). On

M. León (✉)
Institute of Humanities, CONICET-National University of Córdoba, Córdoba, Argentina
e-mail: malena.leon@mi.unc.edu.ar

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_7

culture, it is typically assumed that if something has extraordinary characteristics, it must be the result of the cognitive abilities of a brilliant mind. In discussion with that, in this article, I want to explore how the involvement of Darwinian processes in the cultural domain applies to this modified version of Paley's argument. Particularly, the scope of my argument will consider a paradigmatic case of designed items: creative products. In that sense, I set myself to link two areas of research that have remained, from my perspective, less connected than they should have been: theories of creativity and cultural evolution.[1]

Theories of cultural evolution claim that culture evolves according to Darwinian principles (Acerbi and Mesoudi 2015). Although there are significant differences between schools, a significant subset of theorists – whom Sterelny (2017) calls the Californian school[2] – consider that, in some significant respects, both genetic and cultural transmissions behave in the same way (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985), through a mechanism analogous to natural selection. These theories hold mechanisms of cultural selection guided by different biases (Richerson and Boyd 2005; Mesoudi 2011). In many cases, these theories draw an explicit analogy between natural and cultural selection. Particularly, it has been pointed out that biassed transmission processes – processes by which some cultural variants are favoured over others during the process of cultural transmission – are selective retention processes (Richerson and Boyd 2005, p. 79). In the same vein, Mesoudi defines cultural selection as 'any condition where one cultural trait is more likely to be acquired and passed on than an alternative cultural trait' (2011, p. 64).[3]

This application of Darwinian processes to the cultural domain would have straightforward consequences for creative products that would render false the cultural version of Paley's argument. Creative products are defined as original and functional items: original means that they are of a different type than what already exists; functional means appropriate, useful or adaptive concerning task constraints

[1] Although some relevant theorists of creativity, such as James (1880), Campbell (1960), and Simonton (1999), adopted an evolutionary perspective on this phenomenon, they did not explore the connection between creativity and cultural evolution that I am trying to propose.

[2] The other principal school is what Sterelny (2017) calls the Parisians. The Parisian school is that of cultural epidemiology, a proposal developed by Sperber (1996) and supported by other anthropologists. According to this school, cultural transmission generally implies a transformation, not a replication. I will not deal with it in this paper.

[3] Memetics also explains the cultural change as an evolutionary process (Dawkins 1976; Blackmore 2000; Dennett 1995, 2017). The term 'meme' designates pieces of culture subject to a Darwinian evolutionary process. According to this theory, culture evolves by differential replication of memes. The proposal I will try to put forward shares with memetics the strong analogy between natural and cultural selection. Likewise, Dennett (2017) argues that this theory has much in common with Californian's theory of cultural evolution. While my suggestion is compatible with memetics, I will draw on the Californian theory, particularly Richerson and Boyd (2005), since they focus more on the human processes that explain why some cultural items are replicated and others are not. In other words, if at least some culture changes in the way this theory states, my argument holds. A more detailed analysis of the link between Memetic Theory and Californian theories is beyond the scope of this paper.

(Boden 1991; Csikszentmihalyi 2014; Kozbelt et al. 2010; Kaufman and Glăveanu 2019; Runco and Jaeger 2012; Simonton 1999; Stein 1953; Sternberg and Lubart 1999). According to theories of cultural evolution, cultural products are the result of the 'collaboration' of innumerable people who may not even know each other nor have a complete understanding of the processes to which they contribute. In this sense, they open the possibility of leaving behind the conjecture that behind every brilliant invention there must be a brilliant mind that designs it. The consequences of applying the cultural-evolutionary framework to creative products, however, have not been sufficiently explored. For example, although theories of creativity have increasingly recognised that, to account for creative processes, it is necessary to posit other factors besides individual cognitive abilities, this research area does not usually consider cultural evolution.

How far-reaching the consequences may be, however, is disputed. Under a popular view of theories of cultural evolution, selectionist models built by Californians who, '*assuming the prior existence* of a certain set of variants, intend to explain the *distribution* of such variants' (Baravalle 2017, p. 295, italics and translation mine).[4] Thus, according to a preliminary view, it seems sensible to think that the selective mechanisms postulated by the theories of cultural evolution could not make a relevant contribution to theories of creativity as they could not explain the creation of new cultural variants. If one looks at the mechanisms of cultural evolution proposed by most theories of cultural evolution, one will see that they are of a selective type. But – this hypothetical argument in favour of the separation between cultural and creative processes would continue – the mechanisms of cultural selection only specify the conditions under which one cultural item or trait is more likely to be acquired and transmitted than another (Mesoudi 2011, p. 64). In short, cultural selection would consist of a set of mechanisms that aims at explaining the preferential selection of cultural items but not their emergence. Their emergence could still be accounted for in terms of individual design: creative processes would still be totally prior phenomena that generally occur at an individual scale and are determined by an agent's plan.

In this paper, I will provide an argument against this view and to counter the idea that individual cognitive processes are the only relevant factors to account for the emergence of creative products. Specifically, I will argue that the selective mechanisms postulated by theories of cultural evolution may make a hitherto ignored contribution to theories of human creativity. Specifically, it is a model according to which the two distinctive features of the aforementioned creative products, *originality* and *functionality*, can be explained, at least partially, by resorting to population selective mechanisms. As Boyd et al. (2013) point out in the epigraph to

[4] Although this quotation illustrates the thesis that I intend to dispute, it does not mean that my proposal contradicts what Baravalle defends in this article. Instead, I will also understand biases as mechanisms that modify the frequency of cultural items. However, I intend to argue that such a mechanism can be attributed to an explanatory role different from that assumed at first sight.

this chapter, sometimes it is not individuals who create sophisticated artefacts but rather populations.

To develop my argument, I draw on a dispute that has taken place in evolutionary biology and philosophy of biology: the debate on whether natural selection is a creative force (Wallace 1867; Weismann 1896; de Vries 1906, 1909; Morgan 1909, 1925; Chetverikov 1926; Dobzhansky 1974; Fisher 1958; King 1972; Gould 1977, 1982, 2002; Nei 2013; Orr 2005; Ayala 2007; Razeto-Barry and Frick 2011; Beatty 2016, 2019). The discussion is about whether natural selection constitutes a positive force or merely a negative one. If the latter is the case, selection would only eliminate unfit variants while the onset and direction of evolutionary change would be determined by the production of variation (for example, by mutation).[5] Those who endorse this view understand evolution by natural selection as a two-step process: first, variation occurs and, then, selection takes place.[6] Comparing selection to a sieve would summarise this view (see, for example, Sober 1984, p. 159). On the contrary, if natural selection is a positive force, then it may initiate evolutionary change and impart the direction of evolutionary change.[7] If it was shown that natural selection not only eliminates what is useless but also defines the course and timing of evolutionary change, it would seem that such a mechanism actively contributes to the creation of the designed traits of living organisms. Therefore, the correct analogy would not be that selection is like a sieve, but rather that selection is like a designer.

I take as sound the arguments for the creative force of selection in the natural domain and extend them to the cultural domain. My argumentative strategy consists of extrapolating the arguments in favour of the creativity of natural selection and weighing its applicability to the field of culture by considering cultural selection as a creative force. I then test this thesis by analysing real-world examples that show that the *originality* and *functionality* of creative products can and should be explained by resorting to cultural-evolutionary processes – granting a relevant role to the intervention of mechanisms occurring at the population level. These arguments show that, just as the contention that natural selection is creative disagrees with the conception of evolution by natural selection as a two-stage process, the contention claiming that cultural selection is creative disagrees with the separation between

[5] How this variation comes about is irrelevant for our purposes. I mention mutation because it is the prevailing way variation has been considered to occur. However, variation can also occur by genetic recombination. Moreover, since this discussion of the creativity of natural selection predates the incorporation of Mendelian genetics into Darwinism, to speak of mutation would be anachronistic. For this reason, we will keep the more general term 'variation'.

[6] Although this way of presenting the evolutionary process is quite widespread, it is nothing more than a didactic simplification (Dawkins 1996), which does not adequately describe how the vast majority of evolutionary biologists understand the process to occur (Beatty 2019). Thus, the view of evolution as a two-step process is called into question by the interpretation of natural selection as a creative force, which still prevails within evolutionary biology (Beatty 2019). The presentation of evolution as a two-step process (first variation, then selection) is found, for example, in Mayr (2004).

[7] As we will see later, many notable evolutionary biologists have subscribed to this interpretation of the creative nature of the Darwinian theory.

evolutionary processes and creative processes. Thus, in the same way that natural selection can account for the design of natural objects, so does cultural selection.[8]

I believe these arguments are relevant for constructing a naturalistic model of creativity.[9] While I do not want to deny the relevance of cognitive processes operating at the individual level for explaining the emergence of creative products, I do want to counter the idea that individual cognitive processes are the only relevant factors. In that sense, I favour a pluralistic model in which explanations from different levels can be seen as complementary to each other.

The structure of the paper is as follows: in the first section, I present a characterisation of creative products and evolutionary products according to the relevant literature in each case. Furthermore, I argue that it is possible to establish some equivalences between their distinctive features. In the second section, I expound on Beatty's (2016, 2019) definition of natural selection, according to which it is creative given that it (1) initiates and (2) directs evolutionary change. In the third section, I argue that cultural selection sometimes does indeed operate creatively.

2 Creative Products and Evolutionary Products

Psychologists and philosophers specialised in creativity agree that creative products must be, on the one hand, original or new and, on the other hand, useful, valuable, or functional (Boden 1991; Csikszentmihalyi 2014; Kozbelt et al. 2010; Kaufman and Glăveanu 2019; Runco and Jaeger 2012; Simonton 1999; Stein 1953; Sternberg and Lubart 1999). First, no one would consider a product creative if it is an exact replica of one that already exists. *Originality* is a fairly obvious requirement for creativity. Second, *functionality* is what allows us to distinguish a creative product from a delusion or an idea that is original but has no value whatsoever. On this second

[8] Someone might object that my argument incurs a sort of 'categorial error' since the analogised mechanisms (natural selection and cultural selection) are very different from each other. What I am going to offer is an argument by analogy. Arguments by analogy require that the analogised elements are not identical but share relevant aspects (Gensler 2003). The vast majority of the literature on cultural evolution understands that the processes of cultural evolution and biological evolution differ in important respects (e.g. in culture, inheritance is blending, and there is a lot of directed variation) but that they share relevant aspects. Centrally it is understood that there is, in both cases, a process of selection given that the conditions of variation, inheritance and differential fitness are met (Mesoudi 2011). The other starting point of the argument is that, as some evolutionary biologists have argued, the proper way to interpret how these features operate in the case of natural evolution is to indicate that natural selection sometimes initiates and directs evolutionary change and is, therefore, creative. I intend to argue that, in the same way, cultural selection sometimes initiates and directs evolutionary change (and it is, therefore, creative). Now, for this extrapolation to be well-founded, as we will see below, the significant similarities between natural selection and cultural selection must be conceived as high-level similarities.

[9] As it is known, naturalism has various meanings. Here I am referring to a model that considers scientific knowledge, especially the Darwinian theory of the evolution of species.

criterion, Simonton (1999) points out that its standard of application varies with the sort of domain considered. According to this author, in the technological domain[10] creative artefacts are functional if they work properly.[11] This definition of creativity in which a product to be creative must be both original and functional has been called the *standard definition of creativity* (Runco and Jaeger 2012).

The argument I am going to propose depends on this predominant conception of creativity. However, it is not the only possible conception of creativity. Some believe that a more restrictive definition of creativity is necessary. For example, according to Gaut (2010), creative products should also be the result of an intentional agency exhibiting relevant purposes and understanding. Specifically, Gaut (2010) holds that creative processes are, by definition, those that exhibit relevant purposes, understanding, some degree of judgement, and an evaluative capacity directed at the task at hand (Gaut 2010, pp. 1040–1041). To synthesise these traits, Gaut argues that creative actions must show *aptitude*.

If Gaut's characterization of creative products were adequate, this philosophical endeavour would be in trouble. In other words, if this conception were adopted, the evolutionary-cultural processes I want to consider would not furnish a relevant contribution to the understanding of creative processes. That is because, as we shall see, evolutionary-cultural processes have a population character, involving both cognitive and non-cognitive abilities, and intelligence is distributed in those processes. Therefore, these population processes could be left out of Gaut's proposed aptitude requirement for creativity. However, Gaut's position is in the minority, since the literature on creativity generally assumes that a product must be original and functional to be creative.[12]

Nonetheless, as Kozbelt et al. (2010), as well as Kaufman and Glăveanu (2019) state, the vast majority of theories of creativity address the phenomenon as an individual capacity.[13] Thus, most theories explain *originality* and *functionality*

[10] In the realm of philosophy of technique, some authors have proposed to distinguish between the generic term 'technique' and the more specific 'technology' – reserved for scientifically based industrial techniques – (see Quintanilla 2017, p. 46). However, I will keep the latter as a generic term because it is used as such in the literature on creativity (see Simonton 1999; Kozbelt et al. 2010; Kaufman and Glăveanu 2019).

[11] Simonton's characterisation is consistent with memetic approaches to artefacts but also more generally with any non-intentionalist views of artefactual function, such as those of Dennett (1990, 1995), Kelemen and Carey (2007), and Vermaas et al. (2013). Non-intentionalist perspectives on the artefactual function object to the idea that the designer's intentions determine the artefact's function. In turn, Simonton's characterisation is inconsistent with intentionalist views of artefactual function such as those of Dipert (1993), Kroes and Meijers (2006), or Cuevas-Badallo (2008). Those who rely on the analogy between artefacts and organisms try to move away from the intentionalist perspective.

[12] For a specific defence that the definition of creativity must depend on these products characteristics, see Briskman (1980). For an argument against the idea that creative processes must involve understanding, see Dennett (2001).

[13] Yet, researchers have recently argued that some creative processes are essentially collective (Glăveanu 2011). For example, Csikszentmihalyi's (2014) Systemic Model Theory. But, this theory

appealing to the cognitive abilities of the individual creator (cf. Kaufman and Glăveanu 2019, p. 34). Many of these theories hold or assume that creative achievements were caused by the individual cognitive abilities of some brilliant mind.[14]

Evolutionary theories, for their part, attempt to explain two main aspects: on the one hand, the *diversity* of traits present in living organisms; and, on the other hand, the complexity of the *adaptations* of these traits to their environments (Mesoudi 2011).[15] It is well known that evolutionary theory was originally developed to explain the origin and diversity of biological species, and later extrapolated to the realm of culture. Thus, cultural evolution is a growing scientific field that attempts to provide a naturalistic and quantitative explanation of cultural change. Cultural evolution studies assume that culture evolves according to Darwinian principles (Acerbi and Mesoudi 2015). Thus, cultural evolution studies have proposed a strong analogy between, on the one hand, the *diversity* present in living organisms and cultural items, and, on the other, the *adaptability* of living organisms and cultural items (Mesoudi 2011). Therefore, it would make sense to provide evolutionary explanations for the change of cultural items over time. Some have objected to this approach, since *diversity* and *adaptability* would be explained by the human understanding involved in the design process. However, cultural evolutionary theorists have argued that such understanding is insufficient (Richerson and Boyd 2005; Sterelny 2006; Dennett 2006, 2017). According to Dennett (2017, p. 75), 'top-down design is in fact responsible for much less of the design in our world than is commonly appreciated'. As I pointed out in the Introduction, according to cultural evolution theories, cultural products are the result of the 'collaboration' of countless people who may not know each other nor have a full understanding of the processes to which they contribute (Sober 1994; Sterelny 2006; Boyd et al. 2013). In Sterelny's terms (2006):

> [t]his impressive fit between aboriginal technical, ecological and social organization and their environment is prima facie support for a broadly evolutionary view of culture. For we can safely assume that these adaptive features of their social life were not consciously designed for Australian conditions by some local Plato. Rather, they were assembled piecemeal, just as the biological adaptations of eucalyptus to the same environment were assembled piecemeal (p. 146).

---

still claims that the background and training of creative people are essential for the emergence of creative achievements.

[14] Note that this is the same assumption that the intelligent design argument had (Paley 1802) and it is now used to argue that the world was intentionally created by an intelligent being (see Sober 2019 for analysis).

[15] The idea that these are the main aspects explained by an evolutionary theory is present in the vast majority of evolutionary theories and models – such as Mesoudi's (2011), mentioned above. Although it is convenient for our purpose, it is a non-exhaustive characterisation which places special emphasis on explanations that resort to the action of natural or cultural selection. However, Darwin's theory of evolution also allows us to construct other types of explanations. For example, homologies are explained by appealing to the common ancestor (Blanco et al. 2020).

In developing their theories, Californian theorists of cultural evolution have constructed formal models to explain the distribution of traits over time. To build these formal models, they applied a similar method to that developed by Fisher, Haldane, and Wright for population genetics models (Cavalli-Sforza and Feldman 1981; Richerson and Boyd 2005; Mesoudi 2011). These models focus their attention on the long-term dynamics of cultural variants and not on individuals (Baravalle 2017). Generally, evolutionary theories explain the *diversity* and *adaptability* of traits of organisms or cultural items at a population level.

Cultural selection mechanisms occupy a central place in the theories of the Californians (Houkes 2012). Cultural evolution mechanisms refer to 'any condition where one cultural trait is more likely to be acquired and passed on than an alternative cultural trait (or no trait at all)' (Mesoudi 2011, p. 64). Among the most important processes are the action-driven processes of *content biases*, *frequency-dependent biases*, and *model-based biases* (Richerson and Boyd 2005). *Content biases* refer to the preferential adoption of features based on their intrinsic attractiveness. *Frequency-dependent biases* refer to the preferential adoption of traits by their frequency (e.g. adopting the most popular trait). Finally, *model-based biases* refer to the preferential adoption of traits based on the characteristics of the trait bearer (e.g. if s/he is successful).

Californian theories also account for the mechanisms responsible for providing cultural variation. According to the Californians, the mechanisms of variation are *cultural mutation* and *guided variation* (Mesoudi 2011). *Cultural mutation* refers to 'effects due to random individual-level processes, such as misremembering an item of culture' (Richerson and Boyd 2005, p. 69). *Guided variation* refers to nonrandom changes in cultural variants (see Richerson and Boyd 2005; Mesoudi 2011). According to Richerson and Boyd (2005, p. 69), it is the force that 'results from transformations during social learning, or the learning, invention or adaptative modification of cultural variants'.

The existence of *guided variation* in culture is connected to one of the main objections against theories of cultural evolution. It is objected that the project of explaining cultural change according to Darwin's theory is somehow erroneous, given that one of its principles is that variation is blind. This does not seem to be the case in the cultural domain, where the appearance of new variants is sometimes conditioned by the possibilities of their subsequent selection. So, there is a disanalogy between biological and cultural evolution since in the former, variation is random and, in the latter, variation is directed. However, different researchers have argued that this is not a problem for theories of cultural evolution (Sober 1994, p. 487; Ginnobili 2016; Mesoudi 2011). According to such an objection, in Darwin's theory, variation *has to be* blind to evolutionary change. However, as Ginnobili (2016) points out, in Darwin's theory, variation *can be* blind to evolutionary change. In other words, Darwin claimed that variation does not necessarily have to arise in response to a need. This does not imply that if a characteristic is not 'blind', there is no point in appealing to an explanation by natural selection. Mesoudi (2011) defends something similar when he states that guided variation is not necessary to explain the diversity observed in the natural world, but this does not imply that it is incompatible

with Darwinian evolutionary theory. Thus, there would be no impossibility, in principle, of including guided variation in Darwinian models.

Now, I think that the component of (creative) *originality* is equivalent to that of (evolutionary) *diversity* and that the same is true of (creative) *functionality* and (evolutionary) *adaptability*. *Originality* and *diversity* would be equivalent, for they play the same role in the phenomena at stake (creativity and evolution). Given a set of items, the emergence of a new one could be either a repetition of what already exists (which would imply that a new specimen would come into existence, but not a new type of item) or an innovation. Both, creativity and evolutionary change require the emergence of something different: a novelty in type. The origin of something different allows creative *innovation* and results in the *diversity* on which selection acts to bring about evolutionary change.

On the other hand, (creative) *functionality* and (evolutionary) *adaptability* seem to be the same kind of normative dimensions.[16] *Functionality* is what makes a creative product valuable, worth preserving, and useful. In evolutionary theory, the *adaptation* of a biological or cultural trait refers to its adjustment to the environment. Such adjustment presumably favoured its conservation. Perhaps the sphere of creativity in which this equivalence is most clear is technology. As I pointed out above, in technological creativity, the second component of creative achievement (*functionality*) consists precisely in the fact that an artefact, in addition to being new, happens to work. This equivalence is supported by the comparison of artificial designs with biological organisms, which gave support to the Argument from Design (Paley 1802). Some interpretations argue that Darwinian theory shares Paley's intuition (natural organisms have many 'designed' features), but challenge the theological explanation of that design (Dennett 1995; Gould 2002). This equivalence allows Dennett (1995) to point out that we should literally understand biological organisms as designs. The equivalence also permits that the name chosen by Simonton (1999) for the second component of creative achievements is adaptability. Thus, although not without discussion,[17] we can say that, at least in the technological sphere, the analogy between functioning and adaptability is closer than in the artistic or scientific sphere (because, in the technological sphere, it is easier to determine if a product 'works').

More generally, the extensive literature on creativity assuming the adequacy, in some sense, of a Darwinian account of phenomena we call creative (Campbell 1960; Simonton 1999), also supports equivalences between (creative) *originality* and (evolutionary) *diversity,* and between (creative) *functionality* and (evolutionary)

[16] In fact, Simonton (1999) uses the term 'adaptability' to refer to this dimension of creative products.

[17] There are some ways to complexify what it means for something to work in the technological sphere. For example, Lemonnier (2013) argues that the 'technological choices' that different societies make are more the result of cultural values and social relations than the inherent benefits of technologies themselves. Even if this were true, it seems adequate to assert that, for some technological artefacts, it is possible to determine whether they perform a given function. Therefore, I do not consider that this type of complexity would jeopardise my analysis.

variability. Thus, based on the above equivalences and given that evolutionary theories assume that *diversity* and *adaptation* are, at least in part, explained by population selective processes, it seems sensible to take this debate about the creativity of natural selection as an indicator that these kinds of population processes could be extrapolated to the field of human creativity. The latter would allow us to partly explain *novelty* and *functionality* by the intervention of selective processes of the same population character. Although I consider that this analysis could also be applied to artistic and scientific spheres, for the time being, I will only keep in mind cases of technological creativity, a field in which I consider that the second equivalence (between *functionality* and *adaptability*) is more evident.

In summary, the mechanisms postulated by cultural evolution should be relevant to theories of creativity. The only reason for not contemplating this option would be that evolutionary processes take diverse and adaptive products for granted and only explain their distribution. Some evolutionary biologists have understood evolution in biology in this way. However, this is a minority position and one that can be considered anti-Darwinian (Razeto-Barry and Frick 2011; Beatty 2016, 2019). At stake is the discussion of whether natural selection constitutes a creative force. I present this discussion in the next section.

3 Is Natural Selection Creative?

Many researchers consider that the creativity of the natural selection thesis (hereafter CNST) is at the heart of Darwinian thought (Gould 1977, 1982, 2002; Ayala 2007; Razeto-Barry and Frick 2011; Beatty 2016, 2019). They also point out that some of the considerable objections that the theory has received are linked to this thesis (Gould 1977, 2002; Beatty 2016). But what precisely does the CNST consist of? While there is some polysemy in this thesis,[18] it mainly concerns the relative evolutionary contribution of natural selection and variation. Specifically, those who defend the CNST understand that variation is always available and that natural selection initiates and directs evolutionary change. In contrast, opponents of the thesis argue that variation is not always available and, therefore, natural selection must 'wait' for variation before it can act. In other words, there is a debate about

[18]The CNST also addresses the more general point that natural selection is a mechanism that *produces* evolutionary change and not one that *prevents* it. Thus, Gould (2002) points out that it distinguishes Darwin's theory from earlier theories that also postulated the existence of natural selection. For example, Blyth's theory (1835) held that natural selection is a process that eliminates extreme and maladaptive variants and thereby helps species to retain their essential traits. No one with an evolutionary perspective would argue against the idea that natural selection is a mechanism that contributes to change rather than conservation. Instead, the interesting discussion is about the role of this mechanism in evolutionary change. Thus, we will leave aside this interpretation of CNST, according to which it serves to distinguish an approach in which selection produces change from one in which it prevents change.

whether or not the origin of the traits of organisms is, among others, an *explanandum* of the theory of evolution by natural selection (Razeto-Barry and Frick 2011).

Through a historical-philosophical analysis, John Beatty (2016, 2019)[19] has reconstructed the arguments used at different historical moments by defenders and detractors of CNST. According to his analysis, two main positions can be established, an orthodox one in its defence of the Darwinian theory of the evolution of species and CNST, and a more critical one. In addition to Darwin himself, CNST was supported by Wallace (1867), Weismann (1896), Chetverikov (1926), Dobzhansky (1937, 1974), and Fisher (1958), among others. Some of its principal critics were de Vries (1906, 1909), Morgan (1909, 1925), King (1972), and Nei (2013). Beatty (2016, 2019) calls the position that opposes the CNST 'mutationist'[20] because it claims that mutation initiates and drives evolutionary change. Without ignoring that the mutationist current designates a more specific movement in biology, for the sake of simplicity, I will call the position defending CNST 'selectionist' and the position that criticises it 'mutationist'.

According to Beatty (2016, 2019), the CNST should be understood in terms of (1) natural selection *initiating* evolutionary change, and (2) natural selection *directing* evolutionary change. I will examine each of these assumptions.[21]

The first assumption is that (1) natural selection *initiates* evolutionary change. Although according to Beatty (2016, 2019), this assumption acquires different specificities in different historical moments, I will try to capture its more general meaning. A sensible way to illustrate this is to bring up two scenarios proposed by Darwin in the first edition of *The Origin of Species* (one of which he later eliminated). According to the first scenario, there is a population of wolves whose individuals are very diverse in terms of size and speed. Evolutionary change begins when an environmental change occurs and decreases the number of preys. This change makes the fastest wolves, for instance, those with the largest legs, more fit. After many generations, the entire population of wolves will have longer legs than at the initial time, when there still was greater diversity regarding this trait. As we know, natural selection refers to the non-random differential reproduction of phenotypes within a population. In turn, environmental conditions both favour and hinder the reproductive possibilities of living organisms. Thus, the environment

[19] Beatty (2016, 2019) makes a detailed historical reconstruction of the different instances in which the debate about the creativity of natural selection took place and the positions at stake. I will not go into that grain of detail here since it is irrelevant to my analysis.

[20] According to Beatty (2016, 2019), this position was held by mutationists such as Hugo de Vries and Thomas Hunt Morgan, and by neutralists such as Jack King and Thomas Jukes and Daniel Hartl and Clifford Taubes. However, they all share the idea that the TCSN is false. For this reason, he calls these neutralists 'neo-mutationists'.

[21] Beatty (2016) is offering a reformulation of Gould's (2002) proposal. According to Gould, CNST rests on three assumptions regarding the production of variation. First, variation is abundant and takes place in all directions. Second, while large-scale variation can occur, small-scale variation serves as the material for evolutionary change. Finally, the production of such variations is 'decoupled' from the direction of evolution. For the sake of length, I will directly consider Beatty's (2016) proposal, which I judge to be superior.

constitutes *selective* evolutionary pressures insofar as it impacts the reproductive success of phenotypes in the population. Therefore, an environmental change that modifies the reproductive success of individuals, decreasing the reproductive success of some of them, implies the natural selection of specific phenotypes. Hence, according to this scenario, natural selection initiates evolutionary change.

In the early editions of his famous book, Darwin proposed a second scenario which was, though, later eliminated. According to it, another variable triggers evolutionary change (let us call it a mutation). In this case, there appears a variation previously unavailable in the wolf population (such as a new food preference). This variation confers a high survival value, so natural selection preserves it and, many generations later, it is present in the entire population. Beatty (2016) points out that Darwin is satisfied only with the first conjectured scenario, which gives selection, rather than mutation, a more significant role.

Beatty's second assumption is that natural selection is creative to the extent that (2) it *directs* evolutionary change, 'for example by *creating the variation that it subsequently acts upon*' (Beatty 2019, p. 705). This principle has to do with the cumulative character of selection. The discussion in biology has been expressed in a simplified way as follows. In relation to the *mutationist* view, if natural selection was the only force involved, species might change up to a certain point, but then evolutionary change would come to a halt. According to this position, natural selection eliminates genetic variation to the point where evolution stops and then the appearance of new beneficial variations is necessary for the evolutionary change to restart. Thus, evolution would 'consume its own fuel' (Gould 2002, p. 142). In contrast, proponents of the CNST argue that beneficial variation is always present. This means that evolution by natural selection never stops due to lack of variation; and the process is initiated, directed, and redirected entirely by fitness differences in genes or phenotypic traits and fitness changes in fluctuating environments. Thus, the action of natural selection defines the direction of evolutionary change.

Assumption (2) refers, at least for early *selectionists* such as Wallace (1867) and Weismann (1896), to whether selection can shift the *range* of variation. This would occur if selection in a particular direction results in the production of subsequent variation in the same direction. In other words, the discussion at stake is whether selection, when choosing one variable within a range of possibilities, can make the possibilities available in subsequent generations 'move' in that direction.[22] Thus, the question about the responsibility of natural selection for the variation on which it then acts is not only about an increase in the proportion of an advantageous trait. Instead, as evolution by natural selection moves in a particular direction, there is an increasing amount of variation in that direction for natural selection to continue to

[22] However, as a reviewer remarked, this is not incompatible with asserting the importance of mutation within the evolutionary process. Indeed, *selectionists* recognise that mutation is ever present and in all directions (Beatty 2016).

act on.[23] According to CNST advocates, this is the case: natural selection shifts the range of variation. In contrast, opponents of CNST assume that if it were true that natural selection shifts the range of variation on which it then acts, then CNST should be considered true; but, according to *mutationists*, 'evolution by natural selection doesn't work like this' (Beatty 2016, p. 673).

However, evolutionary biologists associated with the Modern Synthesis offered a reformulation of assumption (2) that natural selection *directs* evolutionary change (Beatty 2019). While previous proponents of the CNST believed that the most appropriate level of variation for evolution was individual genes, these scientists consider that variation occurs at the level of genetic combinations. Selection leads to the emergence of new successful combinations of genes. While genetic mutations are a matter of chance, new gene combinations are, to some extent, the product of natural selection, as this mechanism would have preserved, in the past, some of the components that would later form part of promising combinations.

Some members of the Modern Synthesis went further and pointed out that evolution by natural selection actively favours the accumulation of genetic variation (e.g. Chetverikov 1926; Dobzhansky 1937). Without going into technical details, suffice it to say that certain biological processes allow us to conceive of species as 'sponges' that accumulate more genetic material than what is phenotypically selected (Dobzhansky 1937).[24] Think, for example, of undetectable variability in the form of unexpressed recessive alleles. It is a genetic material that is not making any selectable phenotypic contribution, but can serve as 'raw material' for future mutations. Thus, according to the scientists of the Modern Synthesis, natural selection actively favours the accumulation of genetic variation.

Razeto-Barry and Frick's (2011) reconstruction of CNST refers only to Beatty's second assumption (2). Their way of presenting this thesis may be illuminating. In their terms, natural selection is a creative force because it 'allows' adaptations of a high degree of complexity to emerge, which, in probabilistic terms, would be very

[23] The following quote from Beatty (2016) on how selection in Darwin's theory can imprint a particular direction on subsequent variations may be illustrative: 'By selective "accumulation," he did not just mean increasing the proportion of an advantageous trait within a species, as for example when an ancestral flying squirrel is born with a flap of skin, between its fore- and hind flanks, that is larger (say x+) than the flap possessed by other members of its species (say x), and the initially rare x + variation becomes more and more common. Rather, he was referring to the way in which selection in favour of larger flaps increases the mean flap volume from x, to x+, to x++, to x+++, etc. And the important point here is that, as evolution by natural selection proceeds in the direction of larger flap volumes, ever larger variations become available for natural selection to act upon' (Beatty 2016, pp. 667–668). Beatty (2016) makes a detailed case for selection being responsible for the variation on which it then acts are not incompatible with the Darwinian principle that variation is random (see Beatty 2016, pp. 662–670).

[24] There are various processes by which natural selection preserves variation (heterozygote advantage, disruptive selection). Dobzhansky groups these processes under the label of *balancing selection*. Thus, *balancing selection* refers to a series of selective processes by which multiple alleles (different versions of a gene) are actively maintained in the gene pool population at frequencies higher than those expected from genetic drift alone.

difficult to appear by the simple action of random mutation. Thus, natural selection 'makes more probable the occurrence of types of sequences of phenotypic steps that seem impossible (in other words, extremely improbable) to occur by the random accumulation of changes' (Razeto-Barry and Frick 2011, p. 6).

In summary, I note that, according to Beatty (2016, 2019), natural selection is creative because (1) it *initiates* evolutionary change and (2) it *directs* evolutionary change (e.g. by creating the variations that it subsequently acts upon). I consider that assumptions (1) and (2), as offered by Beatty's definition of CNST, can be taken as *criteria*, i.e. as rules, that would allow us to determine more or less clearly whether we are dealing with a creative force.[25] In the discussion I have reconstructed, the phenomenon that these criteria identify was whether natural selection is a creative force.

In contrast to the above two assumptions (i.e. against CNST), *mutationists* argue that mutation initiates and drives evolutionary change. Note that the presentation of evolution as a two-stage or two-factor process (first variation and then selection) seems consistent with this *mutationist* perspective. According to this presentation, the evolutionary process consists of a first stage concerning the origin of variations (or mutations), which initiates and directs evolutionary change, and a second stage concerning selection, which must 'wait' for the mutation to act and then simply 'chooses' among the available options.[26]

At the beginning of the chapter, I pointed out that I intend to argue that creative processes and evolutionary-cultural processes are not two separate and successive spheres: first creation, then cultural evolution. Such a conception is analogous to the *mutationist* perspective. On the one hand, the separation between creative processes and evolutionary-cultural processes assumes the following: individual creative processes give rise to original and functional (or diverse and adaptive) cultural products, while evolutionary-cultural processes merely determine which of them persist, which disappear, and which are replicated. On the other hand, the mutationist position understands evolution as a two-step process: first, mutation which results in diverse traits, some of which are also adaptive; and then selection, which discards those non-adaptive traits while conserving and replicating the adaptive ones. Thus, according to both conceptions, there is a first stage (creative processes and mutations) in which the *original* (or *diverse*) and *functional* (or *adaptive*) traits of cultural

[25] This is not to be confused with the definition of creative *products* given in the first section of the chapter. I am proposing that when a *process* 'behaves' in the way that either of these two assumptions indicates, I will consider that we have good reasons to assume that such a *process* is making a relevant contribution to the emergence of an original and functioning *product*. The reasons why I consider that processes that 'behave' in this way can be considered 'creative' will become clear later.

[26] Although the presentation of evolution by natural selection as a two-step process is quite widespread, its literal interpretation opposes CNST. This presentation is a way of expounding the theory to simplify it, which leads to confusion (Dawkins 1996; Beatty 2019). To maintain an orthodox Darwinian position on the creativity of natural selection, it would be desirable to avoid the two-step presentation.

products originate. And, according to both conceptions, there is a second stage (evolutionary-cultural processes and natural selection) that only modifies the resulting frequency of the first stage.

As I have said above, I aim to discuss the conception that creative processes and evolutionary-cultural processes are separate and successive stages. To this end, I will argue that cultural selection can, at least in some cases, be a creative force. To do that, I will draw on Beatty's definition of the creativity of natural selection. As I note, taking Beatty's (2016, 2019) reconstruction, natural selection can be considered a creative force insofar as it (1) initiates evolutionary change and (2) directs evolutionary change. I hold that these criteria can be extrapolated to the realm of culture, thus allowing us to identify whether cultural selection can be a creative force.[27]

Each criterion will indicate a different way in which cultural selection will be playing a creative role. Where these are satisfied, the distinctive aspects of creative products (originality and functionality) will be partially explained as effects of selectionist processes operating at the population level. Thus, these criteria contribute to the pluralistic model I want to advocate. According to my model, the assumption that originality and functionality of creative products are only the effects of cognitive processes operating at the individual level must be set aside. Instead, creative products are the effect of cognitive and non-cognitive processes taking place both at the individual and population levels.

4 Is Cultural Selection Creative?

In the previous section, I argued that we can understand assumptions (1) and (2) as criteria that allow us to detect the creativity of natural selection and I anticipated my intention to extrapolate them to the field of culture. This section will be devoted to the latter task. For this purpose, I will consider each criterion separately. This will allow me to analyse first whether it is possible to strip them of their biological specificity, and then to evaluate whether it is reasonable to characterise a mechanism that satisfies them as creative or not. Then, I will try to propose the extrapolation of these criteria to the field of culture and introduce some examples that satisfy them. In this context, the expression 'attributing creativity' is equivalent to 'considering that it may be playing a relevant role in the process of elaboration of a creative product'.

[27] In the following section, I will consider cases of the creation of cultural items, which are both original – not of the type that already existed – and functional – I will leave aside those that did not work. One might object that not every original and functional cultural product is a good case of a creative product, an attribute that should be reserved for exceptional achievements. However, theories of creativity recognise that there are different degrees of creativity: little-c, mini-c, Pro-C, Big-C (Kaufman and Beghetto 2009). Thus, all creative cultural products (and also those I chose as examples for my analysis) would fall into one of these categories.

4.1 Selection Initiates Evolutionary Change

As I pointed out, the first criterion for defending the thesis states that (1) selection initiates evolutionary change. I have explained this through Beatty's discussion of Darwin's two alternative scenarios in a wolf population (cf. Beatty 2016, pp. 665–667). The criterion is satisfied in the first scenario where a change in the natural environment triggers evolutionary change – in Darwin's example, a change in the availability of food. That environmental change leads to differential reproduction of those wolves with more favourable traits for obtaining food (in this case, those with longer legs). Let us try to extract this criterion from its biological specificity. What characterises this type of scenario (and similar non-biological ones) is that a new environmental pressure (whether it occurs in a natural or cultural setting) triggers evolutionary change. In other words, it is the selection rather than the transformation of the item that initiates evolutionary change.

In contrast, in Darwin's second scenario, the emergence of a new variant – that is significantly more adaptive than the available variants – initiates evolutionary change. In the case of biology, the second scenario refers to the emergence of a particularly beneficial variant. In the case of culture, the second scenario could refer to the emergence of a new invention, which in some sense is superior to currently available technologies, but whose production was not stimulated by any particular need or problem. The second scenario behaves as the presentation of evolution in a two-stage process: first variation, then selection. Thus, criterion (1) is not satisfied in the second scenario; instead, evolutionary change must 'wait' for variation to act.

Consequently, criterion (1) states that selection – and not the emergence of a new advantageous variation – initiates evolutionary change. It seems sensible to argue that criterion (1) is an indicator of creativity. In a scenario that satisfies this criterion (i.e. one in which the same selective process leads to the emergence of something original and functional), it seems justified to conclude that such a process would bear some responsibility for change. More precisely, the selection processes would pose a new problem to be solved, and to choose the solution.[28] Thus, the environment would delimit the framework or direction in which the change has to occur. The shaping of a framework where the change has to take place is not a null contribution. In fact, in the area of creativity, Csikszentmihalyi (2014) has pointed out that in many cases the most relevant contribution of creative discovery is due to a problem

[28] A certain degree of abstraction is necessary to perform this analysis. In more concrete terms, first, a change in the environment generates a problem; second, the action of biases leads to the replication of one of the possible solutions available. In evolutionary terms, all these components can be understood as part of a selective process. This is the same type of abstraction that is present in the analysis according to which Darwin's first scenario satisfies the first criterion. In this scenario, an environmental condition poses a problem (absence of prey). This leads to some wolves being better equipped to survive and reproduce, so that the traits that help them solve the problem spread through the population more rapidly. All these elements would be part of the process of natural selection.

statement, which, in addition, frequently leads to a delimitation of the type of answer to be given to the problem, even if it is in a coarse sense.

Let us now try to extrapolate this criterion to the field of culture. If the environment posed a new challenge and its resolution entailed a cultural change, this criterion would be satisfied. That would happen in a scenario where a change in the environment posed a problematic situation. Good examples would be the depletion of a natural resource used as raw material, or a new technological need caused by another innovation. Some of these problems might not have a clear resolution. Others, however, could involve an evolutionary-cultural change; as a consequence, a new invention would emerge, satisfying the problem in question and spreading the invention through that population over the years. This propagation is possible because, in human societies, good creations tend to be replicated and used by the entire population. Thus, it is not necessary to constantly *reinvent the wheel*.

Thus, in general terms, one could argue that criterion (1) is satisfied when, in a population that has many copies of cultural variant A, the environment poses a new challenge or problem P (e.g. the scarcity of materials to build A), and, through a selective process, variant B spreads through the population, while A decreases.[29] Variant B would fulfill the same functions as A and could respond adequately to P.

The evolutionary mechanisms that refer to the change in item frequency are selective processes. These processes concern the action of *content*, *model*, and *frequency biases*. Thus, in a scenario such as the one I am conjecturing, these biases would act in consonance propagating item B. If B is an alternative technology to another A, the propagation of B would imply a decrease of A.

Now, while the action of selective mechanisms explains why some cultural items expand and others disappear, it is necessary to point out how the modification of the item would occur. In other words, we need to explain how item B would originate. This question allows us to formulate two different variants of criterion (1). While in one case item B would be invented to solve the posed problem (1.a), in the other case, item B would be selected from an already existing item, which would have been created for other purposes or no purpose at all (1.b).

Let us begin by analysing variant (1.a). According to (1.a), selection initiates evolutionary change as the environment poses a new problem and 'selects' a cultural item created to solve it. In this case, whoever invents the item does so to solve a problem. In other words, it is a variation introduced in a directed way. For these reasons, I would say that the emergence of the new item constitutes a case of *guided variation*. As I pointed out in Sect. 1, the incidence of *guided variation* in evolutionary change does not constitute an objection to explaining such change by appealing to Darwinian selective mechanisms.

[29] While this schematic modelling attempts to capture the typical occasion when criterion (1) would be satisfied, strictly speaking, cultural variant A may not exist. While, in most cases, the new technology would replace an old one, in other cases, it would simply be an artefact that did not exist before.

On the other hand, according to variant (1.b), selection initiates evolutionary change, since the environment poses a problem leading to the replication of an item that already existed, although used for other purposes. In scenarios of type (1.b), a new challenge leads to the novel use of a cultural item that was manufactured and used for other purposes or no purpose at all (e.g. items that are by-products of the intentional manufacture of other products). Palaeontologists Stephen Jay Gould and Elisabeth Vrba elaborated the concept of *exaptation* that I will take on to analyse this possibility.

Gould and Vrba (1982) developed the concept of *exaptation* to identify a missing phenomenon in evolutionary biology explanations. According to these palaeontologists, *exapted* traits arose as by-products of other evolutionary processes or as adaptations to other functions and were co-opted for a new function. A famous example of exaptation is that of vertebrate bones, whose original function might have been to serve as a reservoir for calcium and, later, to protect vital organs and increase internal consistency. Eventually, the transition to terrestrial life led them to take on the function of support. Another well-known example is that of bird feathers. Initially, the function of feathers was to maintain body temperature more efficiently. Today, the feathers of the vast majority of birds favour flight because of their aerodynamic properties.

The article by Gould and Vrba (1982) was intended as a critique of evolutionary explanations that overemphasise the role of natural selection. However, some Darwinists dismissed this critical view by arguing that the idea of some traits being selected for another or no reason and then co-opted for new uses was already present in Darwin's theory (Dennett 1995). In that sense, I consider that the concepts of *exaptation* and *adaptation* are compatible within the same explanatory framework. Thus, we can say that the cases that satisfy criterion (1.b) are those in which evolutionary processes select an *exapted* cultural item.

To sum up, according to our extrapolation, cultural selection may be performing creatively as long as a new pressure from the environment initiates an evolutionary change that either (1.a) selects a variant invented to solve the problem (*guided variation*), or (1.b) selects a variant that has been produced for other purposes (*cultural exaptation*). Next, I will analyse some examples of technological change that will show that these situations have indeed occurred in the history of cultural change.

4.1.1 Some Examples

To present an example that satisfies criterion (1.a), I will turn to Basalla's (1988) historical reconstruction of technological change. According to Basalla, before the existence of self-acting (or automatic) spinning mule, the spinning mule present in cotton mills required the participation of skilled workers called spinners. Since they played a key role, these workers were in a very good position to demand better wages. After the three-month spinners' strike in England, the factory owners sought help from inventors to develop a device that would allow them to manage without

these workers. And that is when the self-acting machines came into existence (cf. Basalla 1988, pp. 137–141). In this way, a cultural item that had remained stable for a long time (the spinning mule) changed radically, because a new pressure from the environment was produced – in this example, the environmental pressure is a socioeconomic conflict. The emergence of self-acting machines provided an answer to this conflict. This response was unfortunate for the employees but beneficial for the owners of cotton companies. In turn, the acquisition of the artefact provided these that allowed the artefact to expand further.

Once the self-acting machines were invented, this cultural variant spread first across England and then Europe. There was a decrease in the number of non-automatic machines and a progressive increase in the number of self-acting ones – i.e. the variant that allowed the problem of production stoppage to be solved. I consider that this example satisfies criterion (1.a) since the origin of the cultural variant constitutes a case of *guided variation* and its subsequent diffusion can be explained as an effect of the action of different *biases*.[30] I will justify this assertion below.

First, the emergence of the self-acting feature is a case of *guided variation* because it is an item created to solve a specific problem. The problem was that spinners were a scarce labour force that could exert union pressure on employers. For this reason, it was particularly tempting for entrepreneurs to acquire a machine that would make it possible to dispense with spinners. Thus, self-acting machines were a valuable invention in this context. It seems reasonable to conjecture that its chances of being invented were higher than those of an alternative device. For example, the spinning machine would make it possible to dispense with a lower-ranking operator. It would thus be a case of *guided variation* rather than *blind* variation.

Second, we have to consider why the self-acting mule has been replicated. The reason is the way its properties interact with the environment. These properties make the cultural item more advantageous than its alternative variant (the non-automatic spinning mule). These advantages are observable by textile mill owners. We can assume, therefore, that *content bias* is the mechanism responsible for the replication of self-acting machines. The *prestige bias* could also have accelerated the speed of the item's expansion. That would have occurred if, for example, those implementing self-acting mules first had already been the most successful ones (or if the implementation made them successful) and if other entrepreneurs had been sensitive to the formers' success. However, the most relevant bias must have been the *content bias*, which led to the multiplication of a technology that provided an answer to a problem

[30] It is beyond the scope of this research to carry out an in-depth analysis of all the empirical aspects related to the example to determine which of the evolutionary mechanisms proposed by the Californian approach to cultural evolution could have been operating in the production of this cultural change. That would require gathering historical information about the change in the frequency of these cultural items (i.e. the self-spinning machine and the old spinning machines) in the years following the introduction of the invention. However, I can make a more relaxed analysis based on the fact that it is an item created in response to a conflict and then expanded.

for factory owners. Note that this process of expansion of the cultural item is not a minor issue, since thanks to this expansion, the self-acting mule has consequently influenced the development of the spinning machines. If it had not expanded, nobody would know today that such a machine existed; by contrast, it would be a rarity, perhaps it would have rusted in a shed. Therefore, the invention of self-acting machines is a case of criterion (1.a).

On the other hand, criterion (1.b) establishes that selection would behave creatively if it initiated evolutionary change by selecting an *exapted* variant.

The notion of *exaptation*, originally proposed for biology, has been extrapolated to the realm of culture and technology (Lass 1990; Dew et al. 2004; Cattani 2006; Larson et al. 2013; Andriani and Cattani 2016; Ching 2016; Dew and Sarasvathy 2016; Garud et al. 2016). In other words, there has been an 'exaptation of exaptation' (Larson et al. 2013, p. 1). Indeed, it has been argued that the concept is more suitable for the realm of culture than for the realm of biology (Larson et al. 2013).[31]

Cultural exaptation refers to the co-optation for a new function of a product (or by-product) that has originated for other purposes or no purpose at all. For instance, microwave ovens are an *exaptation* of a technology that was originally employed by magnetrons in early radar systems (Osepchuk 1984). The original function of magnetrons was to convert electrical energy into electromagnetic energy. These devices were developed to power radars. However, after observing that a bar of chocolate he was carrying in his pocket had accidentally melted, engineer Percy Spencer discovered this alternative use. Nowadays, microwave ovens are used all over the world to heat food. Another example can be found in drug repositioning, which is a fertile ground for exaptations since many health problems have been solved by exploring the unknown effects of drugs already developed and approved for other purposes. One substance whose consumption further changed its function is gin, as it went from being a drink 'used to alleviate circulatory problems to an intoxicating liquor' (Andriani and Cattani 2016, p. 120). Moreover, there has been research on exaptation in natural languages (Lass 1990; Larson et al. 2013). As linguists have shown, languages evolve and some of their grammatical features become obsolete. While many of these obsolete features subsequently become extinct, others may persist as linguistic 'garbage' for many generations. Sometimes, these features even find a new communicative function, becoming *exaptations*. In sum, the concept of exaptation can be employed in culture as well.

As stated, *cultural exaptation* is a phenomenon compatible with theoretical frameworks that attribute a central place to evolution by natural selection. As a consequence, it can be argued that, although the term '*cultural exaptation*' is not used by Californian theorists, it can be incorporated as another type of mechanism of cultural evolution. More precisely, it is a mechanism of *variation* introduction; that

[31] According to Larson et al. (2013), the term *exaptation* has not become widely used in the biological sciences. They hold that *exaptation* lacks a formal definition that distinguishes it from *adaptation*. However, *exaptation* has been adopted with considerable success in studies of the history of technology. Frequently, technological innovations involve the use of a process or artefact in a new context.

is, it introduces a novelty within the diversity of competing options for a given trait.[32]

As Andriani and Cattani (2016) point out, exaptation is rarely considered in historical reconstructions of the origin of novelty. However, it is possible to recover some cases of *cultural exaptation* that may exemplify the situation postulated in (1. b). Optical fibre, whose genesis was analysed by Cattani (2006), is one of them. It is a small diameter glass-embedded fibre currently used in telecommunications. This technology came to replace electric cables since the latter entail more energy loss and, in addition, are affected by electromagnetic interference, which was very problematic in some implementation circumstances. Thus, optical fibre itself can be seen as a case of *exaptation* as it involves a change in the function of old technologies (embedded glass, used, for example, for pots), which are now used in telecommunications. Furthermore, Cattani (2006) conducted a historiographic study with documents on the transformation of the Corning company, a pioneer in the development of fiber optics, which was previously devoted to the development of glass for special items, such as optics, windshields, and cookware. Consequently, there is also a co-optation of knowledge and technologies used to work with glass in the manufacture of certain artefacts for the production of a new artefact. In conclusion, that knowledge and those technologies can be seen as cases of *exaptation*, too (Andriani and Cattani 2016).

In the example above, the pressure to develop a technology to transmit energy avoiding losses and electromagnetic interference led to the use of another technological development: the embedded glass. Therefore, it is a scenario where selection initiates evolutionary change, and the *exaptation* of an artefact occurs. The clear advantages of the new item are the reasons for its spread, replacing the old technology. Hence, we could think of this as the *content bias* effect.

Another example that could illustrate (1.b) are the disc pans used to cook some Argentine dishes. Originally, these disks were part of ploughing machines and, from time to time, had to be replaced, so they became a waste product. Although there are no formal records, it is quite evident that they began to be used to satisfy some other needs. It was probably the need for a large container suitable for cooking food for many people, during many hours, and in direct contact with fire. The disc fulfilled this function adequately, resulting in an efficient way of cooking, which does not require precise regulation of the flames, and is suitable for simultaneously cooking large quantities of a wide diversity of foods in a relatively easy manner. Consequently, it is not surprising that today this artefact is manufactured and marketed in a

[32] Since this is not a random introduction, but a directed one (co-optation has been carried out for the artefact to perform its new function), it could be considered a sub-type of guided variation. This question is irrelevant for the present analysis since I am interested in distinguishing processes involving cultural exaptation from those involving guided variation without *exaptation* (such as the invention of self-acting machines). It is also true that these delimitations will not always be precise and that many inventions should be considered a mixture of both. However, I believe that this fact does not preclude the possibility of making the distinction.

mass and customised way; that is, it is no longer obtained by ordinarily recycling old plough discs.

Although there are no records on the number of discs[33] present in the Argentine population so far, it is sensible to conjecture the following: discs are a cultural variant that emerged at some point in history (presumably on more than one occasion), then they were replicated by imitation (allowed by horizontal, oblique, and vertical transmission), and, finally, they began to be manufactured and marketed autonomously; so they increased in quantity. Therefore, although the emergence of discs was a response to an environmental challenge, discs were produced by way of *exaptation,* and later extended by a *content bias* mechanism. If it were proven that, after the television appearance of the famous Argentine chef Francis Mallman using a disc, sales increased significantly, the disc would also fulfill the prestige *bias*.

In short, I have presented some examples that satisfy criteria (1.a) and (1.b). In other words, these are scenarios where cultural selection is creative, given that a new pressure from the environment initiates an evolutionary change that either (1.a) selects a variant invented to solve the problem (*guided variation*), or (1.b) selects a variant produced for other purposes (*cultural exaptation*).

4.2 Selection Directs Evolutionary Change

Beatty's (2016, 2019) second criterion notes that natural selection is creative to the extent that (2) it directs evolutionary change. This assumption connects to what Razeto-Barry and Frick (2011) point out about the creativity of natural selection. The authors argue that natural selection allows adaptations of a high degree of complexity to emerge, which would be statistically almost impossible to occur by the action of mutation alone. For his part, Beatty (2016) argues that Darwinists of different times have formulated assumption (2) differently. Thus, Wallace (1867) and Weismann (1896) contended that natural selection directs evolutionary change, to the extent that it changes the range of available variation. If this was the case, selection in a particular direction would result in the production of subsequent variation in the same direction. For their part, the scientists of the Modern Synthesis stated that natural selection actively favours the accumulation of genetic variation since variation is the result of novelty in genetic combinations. Each of the above two variants would allow us to elaborate a different version of assumption (2) of the creativity of cultural selection.

Let us strip this criterion of its biological specificity and analyse whether it is sensible to regard any cultural selection process that fulfils this characteristic as creative. According to this criterion, a selective mechanism would be creative if it directs the variation upon which it acts. Thus, it does not limit itself to merely filtering out the available options. Instead, it is partly responsible for the course that

[33] In this paragraph, we are talking about discs as pans.

will be taken by the available variations on which selections will continue to operate. Hence, the selection process would exhibit active participation, which makes it somewhat responsible for the outcome. Therefore, it can be considered that, if a selective mechanism directs variations, part of the explanation for the *originality* and *functionality* traits of creative products must be provided by appealing to population-type selective processes.

Moreover, in the Modern Synthesis scientists' version of assumption (2), selection directs the course of variation since it actively favours the accumulation of genetic variation (Chetverikov 1926; Dobzhansky 1937). Indeed, it seems sensible to concede creativity if we detect that these selective processes keep variations that are not useful but could be helpful in the future (similar to the Modern Synthesis argument that species accumulate genetic variability). However, this would be a different way of attributing creativity compared to the change in the range of variation, so they should be distinguished from each other.

Consequently, it can be stated that cultural selection will be playing a creative role whenever this mechanism drives evolutionary change; either because (2.a) it simply changes the 'range' of variation, or (2.b) it retains elements that are not useful at a given time but may be helpful in the future.

Criterion (2.a) would be satisfied in a scenario such as the following one. Let us imagine that the action of biases leads to a set of available variants among which those whose values are in an 'extreme' are selected. For example, if there are five options of different sizes, selection mechanisms choose the option with the largest size. It is to be expected that, in subsequent generations,[34] the available variants will have the selected size and, in addition, there will be others of an even larger size.[35] In other words, the 'range' of variations would have shifted.

Second, criterion (2.b) would be satisfied in a scenario in which cultural selection processes preserve some designs that in the future may contribute to the elaboration of new ones. More specifically, one can say that this criterion is satisfied if, on the one hand, not absolutely all existing cultural variants are preserved, and, on the other, not only those that are useful at a given time are preserved either.

4.2.1 Some Examples

Let us begin by analysing criterion (2.a) according to which cultural selection constitutes a creative force since it changes the 'range' of available variation. One

[34] The notion of 'generation' in theories of cultural evolution depends on the context of analysis. That has to do with the fact that some cultural variants are more stable than others. Thus, for cultural variants that tend to remain stable over an individual's lifetime, the cultural generation may coincide with the biological generation; whereas, for cultural variants that tend to change more frequently, the generations will tend to be much shorter.

[35] For example, if in one generation there are size options (i.e. '1', '2', '3', '4', and '5'), and the biases lead to the selection of artefacts of size '5', in the next generation, the available size options will be '4', '5', '6', '7', and '8'.

might think that this occurs more evidently in culture than in biology. For example, consider the change in the weight of cell phones from 1980 to 2000 (Farley 2007). At a certain point, companies decided to develop some lighter-than-average options. Market demand took that direction since many people found it convenient to choose lighter-weight phones. Thus, we can assume that the production of lighter models has been longer than that of heavier ones, leading to a multiplication of the lighter variants and a decrease of the heavier variants – since they are no longer produced, and some of the existing specimens broke down or were scrapped. However, it is relevant that, in addition to this unequal replication of the different available variables, we can think that in the following generations the new models (the new variants) will include increasingly lighter options. Consequently, selection would not simply have chosen some of the available options but it would also have changed the range of available variations. In other words, it is the result of different biases. Presumably, the most relevant is the *content bias* since lighter phones would represent clear advantages over heavier ones. Perhaps, the *prestige bias* may also have been at work if, for example, companies had used the advertising strategy of showing stereotyped images of successful people using lighter phones.

Selection would be exhibiting creativity in a different sense if it met criterion (2.b). According to (2.b), selection is creative if it retains elements that are not useful at a given time but may be helpful in the future. Something similar to this happens with programming codes. Sometimes, a new version of a program leaves an old part of the code useless. Instead of deleting that part of the program, programmers leave it in suspension, in square brackets. In this way, they keep it in case it becomes useful again in the future.[36]

Companies engaged in research and development also serve as an example to meet criterion (2.b). Andriani and Cattani (2016) point out that many firms tend to intentionally retain knowledge, procedures, and designs that at some point have become obsolete but could be reused or exapted in the future. Thus, for example, Hargadon and Sutton (1997) analyse how IDEO, a design and construction company, organises innovation. IDEO initially focused on designing consumer products (from toothbrushes to office furniture to computers). However, in 2001 they began to focus more on 'consumer experiences', designing products such as non-traditional classrooms. Hargadon and Sutton (1997) argue that the reason for this company's innovation was that they had encouraged the storage, retention, and retrieval of knowledge. In fact, the company had retained not only knowledge that was clearly

[36]For example, the programmers of a new version of certain software (e.g. version 7) may add a component that renders a segment of code from the previous version (e.g. version 6) useless. They put that segment in square brackets and 'comment it out'. Eventually, if in the future, the developer of version 8 needs the segment in question again (present in version 6, absent in version 7), he merely has to remove the brackets that rendered it useless (this could happen if, for example, the programmer of 8 decided to remove that element of version 7 that rendered the segment useless). However, the same process of recovering the old segment could be done by obtaining the code of version 6 by another route. For its part, this programmer practice implies a conscious decision to leave behind a design that may be useful in the future.

useful, but also knowledge that had no obvious application. The company had been actively working on the conservation of knowledge that might prove valuable. Andriani and Cattani (2016, p. 125) call this preservation process the conservation of the 'memory of an organization'.

Finally, some developments brought about by the so-called 'maker culture' are also examples of criterion (2.b). This contemporary culture can be regarded as an extension of the so-called 'do-it-yourself' movement, but based centrally on new technology and the use of digital tools. It is mainly interested in engineering-oriented activities, such as electronics, robotics, and 3D printing. Makers are those people who design and produce their own artefacts (Anderson 2012), but they do not do so individually. The 'maker culture' emphasises the potentialities of repeatedly using the 'copy and pastes' strategy for standardised amateur technologies while encouraging the adaptation and reuse of designs published on websites and maker-oriented publications. This movement has encouraged the creation of virtual repositories where different contributors share designs and ideas. Other people use and combine designs made by strangers to build their own artefacts. In some cases, designs are used for the same purpose as were originally intended and, sometimes, for a different purpose. This is a conservation space to store designs that have no obvious use at the moment but may have some in the future.

I have brought up examples of cultural mechanisms responsible for conserving elements that would otherwise have been eliminated. They are, in this way, linked to selection mechanisms, which refer to the conditions by which some cultural items spread and others disappear – rather than to the emergence or modification of an item.

In sum, I have presented some examples that satisfy criteria (2.a) and (2.b). In other words, these are scenarios in which cultural selection is creative, given that it directs evolutionary change; either because (2.a) it simply changes the 'range' of variation, or (2.b) it retains elements that are not useful at a given time, but may be useful in the future.

The examples I have analysed show that, in principle, there are cases in which cultural evolution meets the criteria to be considered a creative force. They reveal different ways in which cultural selection processes, on some occasions, do play a relevant role in the elaboration of creative cultural items. Although future research supported by different types of empirical evidence could complement the present analysis, I consider that what I have offered so far constitutes a satisfactory advance in arguing that cultural selection is a creative force. As a consequence, evolutionary-cultural processes may be relevant to theories of creativity. If evolutionary-cultural processes affect creativity, the view that creative processes are limited to individual cognitive abilities must be abandoned. In the following section, I try to specify in what sense each of these scenarios contributes to this pluralistic view of creative processes.

4.3 Population Processes Matter to Creativity

Let us recapitulate, then, the criteria that I have extrapolated into the cultural domain and for which I found examples. I have argued that cultural selection may be playing a relevant role in the creation of a new cultural item as long as:

(1) a new pressure from the environment initiates an evolutionary change, which either

   (a) selects an invented variant to solve the problem (*guided variation*) or
   (b) selects a variant produced for other purposes (*cultural exaptation*);

(2) it directs evolutionary change; either because it

   (a) changes the range of variation or because
   (b) retains elements that are not useful at a given time but may be useful in the future.

Each of these criteria indicates four different ways in which selective mechanisms would be operating and thereupon contributing to the emergence of creative products. In addition, these criteria would allow us to subtract explanatory weight from the action of the individual mind's cognitive abilities.

Criterion (1.a) identifies cases in which the biases select one of the new variants that arise in response to a problem that was posed by the environment. On the one hand, I am interested in pointing out that the posing of a problem – and not only its resolution – is a considerable contribution to the emergence of a creative product. In fact, some creativity theorists have argued that the elaboration of a problem is one of the most relevant instances of the creative process (Csikszentmihalyi 2014). However, when we attribute creativity to someone, we generally consider the resolution of a problem, but not the posing of a problem. In short, the problem statement has guided the exploration and search for solutions in a particular direction and has contributed, at least modestly, to the emergence of the creative solution.

On the other hand, I am interested in indicating that when a suitable solution to the problem in question appears, cultural selection 'recognises' it, favouring its conservation and replication. In one of the examples previously analysed, the spinners' strike causes the problem and the solution emerging by guided variation is the self-acting machine. The conservation and expansion of self-acting machines can be explained mainly by the action of *content bias*. As this bias refers to the identification of how beneficial a given cultural item may be, we can assume that it involves standard cognitive abilities and that such abilities are 'distributed' in a population. So they are much less grandiloquent than those usually attributed to a single creative mind. Therefore, even if the solution to a problem, in scenarios such as the one analysed, is created by a single mind, many others decide whether it is a good one. In addition, I note that in the last examples and in some of the following ones, the *prestige bias*, whereby individuals imitate the trait that successful people carry, may also play a role. Although this bias involves cognitive skills, such as identifying who is successful in a given field, these are cognitive skills of a very

different type from those generally assumed to be at play in creations. Californians understand that these abilities of our minds are the product of evolution itself, both by natural and cultural selection (Richerson and Boyd 2005). Thus, on the one hand, criterion (1.a) suggests that the problem statement (which is then solved using creative solutions) often does not have an author; however, it is posed by the environment itself. On the other hand, criterion (1.a) points out that, even if the answer to the problem were devised by a single individual using his individual cognitive skills, the recognition that it is a good answer would be in the hands of many other individuals, using standard cognitive skills (such as the recognition of the benefits of an artefact) and other less standard ones (such as the recognition of who is successful). Consequently, the environment would have contributed to the emergence of an original product by presenting a problem, and the standard and non-standard cognitive skills distributed throughout the population would have contributed to the selection of the option that, in addition to being original, worked out.

Second, criterion (1.b) selects those cases in which biases have *exapted* a cultural item to respond to a challenge posed by the environment. On the one hand, again, it is a problem without a definite 'author' and whose solution has been selected by the action of different biases. All the people involved in its replication have either recognised that it was a suitable solution or inferred so indirectly by imitating the most successful ones. Moreover, the exaptation of the artefact involved skills very different from those usually assumed to be involved in creation. In the examples given, the design of the *exapted* artefact fulfills the function assigned to it. However, this is, to some extent, haphazard. I can assume that the creation of the original artefact was guided by intelligent design and that the idea of implementing it for a new function is another intelligent decision. However, this intelligence is distributed in at least two different instances and by two different people, and it requires some dose of good luck. When one starts to dig up some facts, there seem to be many historical examples of cultural exaptations, but our intellectualist way of conceiving history has often made them invisible. I have analysed examples of *exapted* items that constitute creative elements, given that they are original and functional. Although they are items that existed previously, the originality may consist of a new way of using them. As a consequence, the confluence or encounter between an old item and its new 'niche' would guarantee originality. Hence, accidental elements played a relevant role in the emergence of these original and valuable items. For example, a problem was posed by the environment, a pre-existing artefact met this environmental challenge, and the old design and the new function matched. Thus, it is more a matter of taking advantage of a previous design and changing its function. Therefore, the cognitive skills involved in the elaboration of the solution are, at least, distributed into more than one person: the one who makes the original design and the one who employs it in a novel way. Moreover, the final achievement is more linked to chance than to planning.

Third, criterion (2.a) identifies circumstances in which selection directs evolutionary change because it modifies the range of available variations. I have pointed out that this occurs, for example, when users of a product choose to purchase models

with features whose values are at some 'extreme' of the available possibilities. For example, when buying a cell phone, most people pick lighter models. This preference has led developers to design products in which this value is emphasised. Hence users have more influence than designers on the change direction of cell phones design. The proposed example – change in the frequencies of different cell phone models over time, where the lightest ones are the majority – may involve the development of some cell phone models that can be considered creative. In such a case, the design of lighter models involves an original and functional solution to the problem of reducing weight without reducing technological capacity. While the development of these designs surely involves cognitive skills of a group of individuals devoted to design and development to innovate on existing technologies, the framework in which this innovation must take place is determined (as in 1.a) by an external element, in this case: users' behaviour. By choosing certain models over others, users have caused the range of available variations to shift in a particular direction over time. I have conjectured that this shift has been an effect of biases, such as *content bias* and *prestige bias*. Therefore, we are dealing with cognitive skills that are distributed among the population and are different from the single-mind-creating something type typically supposed to be at play.

Finally, criterion (2.b) selects cases in which a design that has no apparent utility yet but may have one in the future is retained. This practice has proven to be very productive, so it is quite common in different fields. The fact that this practice is rewarding shows that, if an expert does not know what a particular design can be used for, it should not be inferred that the design is useless. Although I have not been able to conjecture a typical Californian explanation of this process, I have pointed out that it is a selective type of mechanism, since it consists of conserving something for the future, preventing it from disappearing. The type of cognitive skill involved in this case is very different from the one usually assumed to be at play when someone creates something, since it does not imply knowing something (e.g. that an idea might work), but recognising that a design might have a utility, which has not been identified yet.

In sum, these are some scenarios in which the importance of individual cognitive skills is not absolute. Instead, it seems that mechanisms operating at the population level are significant in the emergence of some creative products.

5 Conclusions

I have argued that there is an equivalence between the distinctive features of creative products and the distinctive features of evolutionary products. Specifically, I have argued that there is an equivalence between *diversity* and *originality* and between *functionality* and *adaptability*. Thus, if evolutionary processes give rise to *varied* and *adaptive* products, one may expect that these same processes would also be relevant in explaining the production of *original* and *functional* outcomes. In other words, those evolutionary-cultural processes are relevant to the theories of creativity.

After that, I discussed the creativity of natural selection within the field of biology. I stated that, according to Beatty (2016, 2019), natural selection is creative because it (1) initiates and (2) directs evolutionary change. I argued that these assumptions can be considered as criteria that allow us to identify whether a mechanism is operating in a creative way or not.

Lastly, I have stripped criteria (1) and (2) of their biological specificity. After extrapolating each of them into the domain of culture, I have presented real-world examples that support our analysis. Thus, I hope to have shown that cultural selection performs creatively on many occasions. I have then made explicit in what sense, in these different instances, the *original* and *functional* character of creative products should be partially explained by processes that take place at a population level and that involve abilities of different kinds, both cognitive and non-cognitive.

In summary, I have provided reasons to conclude that the selective mechanisms postulated by cultural evolutionary theories can play a relevant role in creative processes.

Acknowledgements I am very grateful to Silvia Carolina Scotto, Laura Danón, Andres Alejandro Ilcic, and Nicolás Sánchez for their helpful comments and suggestions. Funding: This research was funded by the research project CONSOLIDAR 33620280100389CB (SECyT UNC, Argentina), by the research project PIP 11220200103107CO (CONICET, Argentina), and by the research project PICT 2020 N° 1653 (ANPCyT, Argentina).

References

Acerbi A, Mesoudi A (2015) If we are all cultural Darwinians what's the fuss about? Clarifying recent disagreements in the field of cultural evolution. Biol Philos 30(4):481–503

Anderson C (2012) Makers: the new industrial revolution. Random House, New York

Andriani P, Cattani G (2016) Exaptation as source of creativity, innovation, and diversity: introduction to the special section. Ind Corp Chang 25(1):115–131

Ayala FJ (2007) Darwin's greatest discovery: design without designer. Proc Natl Acad Sci 104(1): 8567–8573

Baravalle L (2017) El papel del pensamiento poblacional en la teoría de la doble herencia. Sci Stud 15(2):283–305

Basalla G (1988) The evolution of technology. Cambridge University Press, Cambridge

Beatty J (2016) The creativity of natural selection? Part I: Darwin, Darwinism, and the mutationists. J Hist Biol 49(4):659–684

Beatty J (2019) The creativity of natural selection? Part II: the synthesis and since. J Hist Biol 52(4): 705–731

Blackmore S (2000) The meme machine. Oxford Paperbacks, Oxford

Blanco D, Roffé A, Ginnobili S (2020) The key role of underlying theories for scientific explanations. A Darwinian case study. Principia 24(3):617–632. https://doi.org/10.5007/1808-1711.2020v24n3p617

Blyth E (1835) An attempt to classify the 'Varieties' of animals, with observations on the marked seasonal and other changes which naturally take place in various British species, and which do not constitute varieties. Mag Nat Hist 3:40–53

Boden M (1991) The creative mind: myths and mechanisms, 2nd edn. Routledge, London

Boyd R, Richerson P (1985) Culture and the evolutionary process. The University of Chicago Press, Chicago
Boyd R, Richerson P, Henrich J (2013) The cultural evolution of technology: facts and theories. In: Cultural evolution: society, technology, language, and religion, vol 12. MIT Press, Cambridge, MA, pp 119–142
Briskman L (1980) Creative product and creative process in science and art. Inquiry 23(1):83–106
Campbell DT (1960) Blind variation and selective retentions in creative thought as in other knowledge processes. Psychol Rev 67(6):380–400
Cattani G (2006) Technological pre-adaptation, speciation, and emergence of new technologies: how Corning invented and developed fiber optics. Ind Corp Chang 15(2):285–318
Cavalli-Sforza L, Feldman M (1981) Cultural transmission and evolution: a quantitative approach. Princeton University Press, Princeton
Chetverikov S [1926 (1961)] On certain aspects of the evolutionary process from the standpoint of modern genetics. Proc Am Philos Soc 105(2):167–195
Ching K (2016) Exaptation dynamics and entrepreneurial performance: evidence from the Internet video industry. Ind Corp Chang 25(1):181–198
Csikszentmihalyi M (2014) The systems model of creativity. Springer, Dordrecht
Cuevas-Badallo A (2008) Los bioartefactos: viejas realidades que plantean nuevos problemas en la adscripción funcional. Argum Razón Téc 11:71–96
Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
Dawkins R (1996) Climbing mount improbable. Norton, New York
Dennett D (1990) The interpretation of texts, people and other artifacts. Philos Phenomenol Res 50: 177–194
Dennett D (1995) Darwin's dangerous idea: evolution and the meanings of life. Simon and Schuster, New York
Dennett D (2001) In Darwin's wake, where am I? Proceedings and Addresses of the American Philosophical Association 75(2):11–30
Dennett D (2006) From typo to thinko: when evolution graduated to semantic norms. In: Levinson SC (ed) Evolution and culture: a Fyssen foundation symposium. MIT Press, pp 133–145
Dennett D (2017) From bacteria to bach and back: the evolution of minds. W. W. Norton & Company, New York and London
Dew N, Sarasvathy S (2016) Exaptation and niche construction: behavioral insights for an evolutionary theory. Ind Corp Chang 25(1):167–179
Dew N, Sarasvathy S, Venkataraman S (2004) The economic implications of exaptation. J Evol Econ 14(1):69–84
Dipert RR (1993) Artifacts, art works, and agency. Temple University Press, Philadelphia
Dobzhansky T (1937) Genetics and the origin of species. Columbia University Press
Dobzhansky T (1974) Chance and creativity in evolution. In: Ayala F, Dobzhansky T (eds) Studies in the philosophy of biology. University of California Press, Berkeley, pp 307–338
Farley T (2007) The cell-phone revolution. American Heritage of Invention & Technology. https://www.americanheritage.com/content/cell-phone-revolution
Fisher R (1958) The genetical theory of natural selection, 2nd edn. Dover, New York
Garud R, Gehman J, Giuliani A (2016) Technological exaptation: a narrative approach. Ind Corp Chang 25(1):149–166
Gaut B (2010) The philosophy of creativity. Philos Compass 5(12):1034–1046
Gensler HJ (2003) Introduction to logic. Routledge, New York
Ginnobili S (2016) Cultural adaptations: is it conceptually coherent to apply natural selection to cultural evolution? In: Cardillo M, Muscio H (eds) Darwin's legacy: the status of evolutionary archaeology in Argentina. Archaeopress Publishing Ltd, Oxford, pp 1–11
Glăveanu V (2011) How are we creative together? Comparing sociocognitive and sociocultural answers. Theory Psychol 21:473–492
Glaveanu V, Kaufman J (2019) A historical perspective. In: Kaufman J, Sternberg R (eds) The Cambridge handbook of creativity, 2nd edn. Cambridge University Press, Cambridge, pp 9–26

Gould SJ (1977) Ever since Darwin: reflections in natural history. W. W. Norton, New York
Gould SJ (1982) Darwinism and the expansion of evolutionary theory. Science 216(4544):380–387
Gould SJ (2002) The structure of evolutionary theory. Harvard University Press, Cambridge
Gould SJ, Vrba E (1982) Exaptation—a missing term in the science of form. Paleobiology 8(1): 4–15
Hargadon A, Sutton R (1997) Technology brokering and innovation in a product development firm. Adm Sci Q 42(4):716–749
Houkes W (2012) Population thinking and natural selection in dual-inheritance theory. Biol Philos 27(3):401–417
James W (1880) Great men, great thoughts, and the environment. In: Ruse M (ed) Philosophy after Darwin: classic and contemporary readings. Princeton University Press, Princeton, pp 49–55
Kaufman J, Beghetto R (2009) Beyond big and little: the Four C model of creativity. Rev Gen Psychol 13:1–12
Kaufman J, Glăveanu P (2019) A review of creativity theories: what questions are we trying to answer. In: Kaufman J, Sternberg R (eds) The Cambridge handbook of creativity, 2nd edn. Cambridge University Press, Cambridge, pp 27–43
Kelemen D, Carey S (2007) The essence of artifacts: developing the design stance. In: Margolis E, Laurence S (eds) Creations of the mind: theories of artifacts and their representation. Oxford University Press, Oxford, pp 212–230
King JL (1972) The role of mutation in evolution. In: Le Cam ML, Neyman J, Scott EL (eds) Proceedings of the sixth Berkeley symposium on mathematical statistics and probability: Darwinian, neo-Darwinian, and NonDarwinian evolution, vol 5. University of California Press, Berkeley, pp 69–100
Kozbelt A, Beghetto R, Runco M (2010) Theories of creativity. In: Kaufman J, Sternberg R (eds) The Cambridge handbook of creativity. Cambridge University Press, Cambridge, pp 20–47
Kroes P, Meijers A (2006) The dual nature of technical artefacts. Stud Hist Philos Sci 37:1–4
Larson G, Stephens P, Tehrani J, Layton R (2013) Exapting exaptation. Trends Ecol Evol 28(9): 497–498
Lass R (1990) How to do things with junk: exaptation in language evolution. J Linguist 26:79–102
Lemonnier P (2013) Technological choices: transformation in material cultures since the Neolithic. Routledge, London
Mayr E (2004) What makes biology unique? Considerations on the autonomy of a scientific discipline. Cambridge University Press, Cambridge
Mesoudi A (2011) Cultural evolution: how Darwinian theory can explain human culture and synthesize the social sciences. University of Chicago Press, Chicago
Morgan T (1909) For Darwin. Pop Sci Mon 74:367–380
Morgan T (1925) Evolution and genetics. Princeton University Press, Princeton
Nei M (2013) Mutation-driven evolution. Oxford University Press, Oxford
Orr H (2005) The genetic theory of adaptation: a brief history. Nat Rev Genet 6:119–127
Osepchuk JM (1984) A history of microwave heating applications. IEEE Trans Microw Theory Tech 32(9):1200–1224
Paley W (1802) Natural theology, or, evidences of the existence and attributes of the deity, collected from the appearances of nature Rivington
Quintanilla M (2017) Tecnología: un enfoque filosófico y otros ensayos de filosofía de la tecnología. Fondo de Cultura Económica
Razeto-Barry P, Frick R (2011) Probabilistic causation and the explanatory role of natural selection. Stud Hist Philos Biol Biomed Sci 42(3):344–355
Richerson P, Boyd R (2005) Not by genes alone: how culture transformed human evolution. The University of Chicago Press, Chicago
Runco M, Jaeger G (2012) The standard definition of creativity. Creat Res J 24:92–96
Simonton D (1999) Origins of genius: Darwinian perspectives on creativity. Oxford University Press, Oxford

Sober E (1984) The nature of selection: evolutionary theory in philosophical focus. University of Chicago Press, Chicago
Sober E (1994) Models of cultural evolution. Conceptual issues. In: Sober E (ed) Evolutionary biology, 2nd edn. The MIT Press, Massachusetts, pp 477–492
Sober E (2019) The design argument. Cambridge University Press, Cambridge
Sperber D (1996) Explaining culture: a naturalistic approach. Blackwell Publishing, Oxford
Stein M (1953) Creativity and culture. J Psychol 36:311–322
Sterelny K (2006) Memes revisited. Br J Philos Sci 57(1):145–165
Sterelny K (2017) Cultural evolution in California and Paris. Stud Hist Philos Biol Biomed Sci 62: 42–50
Sternberg RJ, Lubart TI (1999) The concept of creativity: prospects and paradigms. In: Handbook of creativity, vol 1. Cambridge University Press, pp 3–15
Vermaas PE, Carrara M, Borgo S, Garbacz P (2013) The design stance and its artefacts. Synthese 190(6):1131–1152
de Vries. (1909). Mutation theory: experiments and observations on the origin of species in the vegetable kingdom. Farmer JB, Darbishire AD (Trans). Open Court, Chicago.
de Vries H (1906) Species and varieties: their origin by mutation, 2nd edn. Open Court, Chicago
Wallace A (1867) Creation by law. Quat J Sci 4(October):471–488
Weismann A (1896) On germinal selection as a source of definite variation. Open Court Publishing Company, Chicago

Incommensurability in Evolutionary Biology: The Extended Evolutionary Synthesis Controversy

Juan Gefaell and Cristian Saborido

Abstract Evolutionary biologists today debate whether it is convenient to revise the standard theory of evolution, or if an Extended Evolutionary Synthesis is necessary. However, the conceptual relationship between the standard theory of evolution (also known as the Modern Synthesis) and a putative Extended Evolutionary Synthesis is not clear. One concept in philosophy of science that has traditionally been put in place to make sense of the conceptual relationship between competing theories or frameworks is that of Kuhnian incommensurability. In a book chapter, Pigliucci argued that the Modern Synthesis and the Extended Evolutionary Synthesis are not incommensurable frameworks and that their relationship is best understood as a business-as-usual extension of our current knowledge about evolution. However, while valuable, we believe that Pigliucci's analysis is limited in several respects. After pointing out what these limitations are, in this chapter we try to provide an alternative analysis of incommensurability between the Modern Synthesis and the Extended Evolutionary Synthesis. We argue that there are compelling reasons to think that both frameworks are incommensurable, thereby leaving the door open for future philosophical explorations.

Keywords Modern Synthesis · Extended Evolutionary Synthesis · Thomas S. Kuhn · Incommensurability · Evolutionary biology

J. Gefaell (✉)
Centro de Investigación Mariña, Departamento de Bioquímica, Genética e Inmunología, Universidade de Vigo, Vigo, Spain
e-mail: gefaell@uvigo.es

C. Saborido
Department of Logic, History and Philosophy of Science, UNED, Madrid, Spain
e-mail: cristian.saborido@fsof.uned.es

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_8

1 Introduction

Current evolutionary biology is immersed in a far-reaching debate about the desirability of revising the standard theory of evolution, also known as the Modern Synthesis (MS) (e.g., Laland et al. 2014; Futuyma 2017; Müller 2017). Some evolutionary biologists claim that many empirical and theoretical advances made during the last decades in diverse branches of biology call for a supposedly 'extended' version of evolutionary theory, which they have accordingly called the Extended Evolutionary Synthesis (EES). Other evolutionary biologists are skeptical, though, of the EES, instead arguing that the MS can perfectly account for these empirical and theoretical advances without revising any of its fundamental tenets.

What is the exact relationship between the MS and the EES? Is the EES just an extension of the MS? Or, conversely, does it represent a more radical break with standard evolutionary theory? These questions are relevant not only to evolutionary biologists themselves but also to philosophers of science interested in how scientific theories change over time (Andersen and Hepburn 2013, online). Scientific change has been a classic topic of study for philosophers of science not only for its historical interest but also because it can give us clues about general questions related to the rationality of science.

One way to interpret scientific change is through the concept of incommensurability. Originally coined by Thomas S. Kuhn and Paul Feyerabend in 1962 (Kuhn 1970[1962], Feyerabend 1962), incommensurability can be defined as the lack of neutral standards from which to compare and assess the merits of competing theoretical frameworks. Incommensurability can be thought of as a proxy for a radical break; if, after analysing two competing theoretical frameworks, enough instances of incommensurability are found, then we can conclude that the transition between these frameworks is not continuous but abrupt. If, conversely, no real signs of incommensurability are detected, then we can argue that the analysed theoretical frameworks are not that different and that there is a significant degree of continuity between them.

In a book chapter, philosopher and evolutionary biologist Massimo Pigliucci has made use of the concept of incommensurability as seen by Kuhn to analyse the purported transition between the MS and the EES (Pigliucci 2017). After examining their conceptual structure, he concluded that the MS and the EES are not incommensurable and that the transition between them is best viewed as a business-as-usual extension rather than a radical break. However, his attempt, while valuable, is limited in some respects. Incommensurability as described in Kuhn's work is a complex concept that, in some cases, can lead to misinterpretations. We suggest that Pigliucci's use of incommensurability includes some of these misinterpretations, thus limiting the validity of his analysis of the MS versus EES controversy.

In this chapter, we try to show why Pigliucci's account is problematic and present an alternative analysis of incommensurability between the MS and the EES. Unlike Pigliucci, we believe that there are substantial reasons to think that the MS and the EES are incommensurable and that the EES does not probably constitute an ordinary

extension of the standard theory of evolution but a significantly different theoretical framework.

But before going any further, we must ask, however, the following question: Does Kuhn still have something meaningful to tell us about how scientific disciplines change? Many scholars would probably doubt if there is any point in employing a Kuhnian concept to analyse a scientific dispute. Today, Kuhn's ideas are often dismissed without due consideration by many philosophers of biology. This is especially the case in the philosophy of evolutionary biology, where Mayr's critique of Kuhn has convinced most philosophers that this author has nothing interesting to offer to understand the historical dynamics of evolutionary biology (Mayr 1994, 2004). However, in recent decades, the Kuhnian concept of incommensurability has experienced a resurge in popularity; many philosophers and historians of science have successfully used this concept to explain certain aspects of scientific disputes throughout the history of science (Wray 2005, 2011; Chang 2013; Politi 2018). In addition, Tanghe and collaborators have recently conducted a Kuhnian analysis of the history of evolutionary biology, concluding that 'interpreting [the] history [of evolutionary biology] through a Kuhnian prism may be somewhat unorthodox but it most definitely makes sense' (Tanghe et al. 2021: 25). Following this recent trend, in this chapter, we would like to show how a distinctively Kuhnian concept (incommensurability), when properly considered, may shed light on how debates in contemporary evolutionary biology unfold.

2 Pigliucci, Kuhn, and Incommensurability

Pigliucci has been one of the first scientists to call for an EES (Pigliucci 2007, 2009; Pigliucci and Müller 2010). However, since his first publications on the subject, he has made clear that this new and purportedly extended theory of evolution does not constitute a radical break with the past (Pigliucci 2007, 2009, 2012, 2017). To argue for this conclusion, in some of these publications this author has made use of Kuhnian concepts. More specifically, in a 2017 book chapter, Pigliucci (2017) used the notion of incommensurability as depicted in *The Structure of Scientific Revolutions* (*SSR*; Kuhn 1970) to argue that the MS and the EES are not incommensurable theoretical frameworks. But what exactly does that mean? To understand Pigliucci's claim, let us first take a brief look at incommensurability in *SSR*.

2.1 Incommensurability in SSR

Kuhn introduced incommensurability in *SSR* to account for the relationship between two competing paradigms that are involved in a scientific revolution (Kuhn 1970; see also Hoyningen-Huene 1990; Sankey 1993; Hoyningen-Huene and Sankey 2001). Kuhn understood paradigms—or *disciplinary matrices*, as he would later

call them—as all-encompassing guides for conducting science that tell scientists how to approach new scientific problems ('exemplars'), which theoretical principles should be employed to explain phenomena ('symbolic generalisations'), how to conceptualise the world ('models'), and which epistemic values a theory should satisfy ('values') (Kuhn 1970). According to Kuhn, these competing paradigms are so different from each other in terms of their exemplars, symbolic generalisations, models, and values that a neutral comparison between them is not possible. The assessment of their scientific merits is always based on the assumptions of one of the two competing paradigms. Kuhn referred to this impossibility of a neutral paradigm comparison as incommensurability.[1]

Many Kuhn scholars have argued that the *SSR* notion of incommensurability covers three different but interrelated domains of scientific practice: methodology, observations, and semantics (see, for instance, Hoyningen-Huene 1993). *Methodological* incommensurability amounts to disagreements over the kinds of problems that each paradigm considers relevant. For Kuhn, different paradigms have significantly different research agendas, up to the point of considering a rival's research agenda as irrelevant, obsolete or, in the most extreme cases, even unscientific (Kuhn 1970). In addition, methodological incommensurability also refers to disagreements over which theoretical or epistemic values should be prioritised when evaluating the scientific quality of a theory or hypothesis. Finally, methodological incommensurability also includes those differences that scientists of competing paradigms have about the ultimate aims of their scientific discipline: Should it aspire to a grand unified theory, or, on the contrary, should it focus on building partial models of phenomena?

Observational incommensurability is the process by which 'proponents of competing paradigms practice their trades in different worlds' (Kuhn 1970: 150). This vague definition of observational incommensurability has led to a great deal of speculation about the correct way of interpreting it. According to the most common interpretation, observational incommensurability would refer to a perceptual process by which observable entities and processes are 'seen' differently by proponents of each paradigm due to the theory-ladenness of observation.[2]

Finally, *semantic* incommensurability entails that scientists from different paradigms employ the same terms in significantly different ways, with different meanings and theoretical implications. For example, according to Kuhn (1970), the meaning and theoretical role of the concept of 'mass' in Newtonian and Einsteinian physics are radically different; when discussing the merits of their paradigms, Newtonian and Einsteinian physicists do not use the term 'mass' in the same way,

[1] In this chapter, we focus only on incommensurability as it is portrayed in *SSR*, not only because this is the notion of incommensurability that Pigliucci employs, but also because we believe it is much better suited to analyse scientific disputes than its taxonomic, post-*SSR* counterpart (for a comparison of both versions of incommensurability as applied to the MS versus EES controversy see Gefaell and Saborido (2022)).

[2] However, as we will see in Sect. 3.3, this is probably not the best way to interpret observational incommensurability.

although they use the same word. Semantic incommensurability causes communication problems between scientists of different paradigms.

In sum, according to Kuhn, incommensurability refers to the supposed absence of neutral grounds to compare competing paradigms at the level of their methodological, observational, and semantic domains.

2.2 What Does Pigliucci Say and Why Is It Problematic?

To establish whether the MS and the EES are incommensurable, in his book chapter, Pigliucci discusses the three aforementioned spheres of incommensurability (methods, observations, and semantics) as applied to this particular controversy, concluding that:

1. There is no methodological incommensurability between the MS and the EES because 'the EES takes on board the same puzzles [i.e., the same research agenda], and the same set of approaches, of the MS, but also adds new puzzles (. . .)' (Pigliucci 2017: 99).
2. There is no observational incommensurability because 'the very same facts that have been catalogued and explained by the MS enter into the empirical corpus of the EES' (Pigliucci 2017: 99).
3. There is no semantic incommensurability because '[k]ey biological concepts (. . .) retain similar and perfectly commensurable meanings' in both theoretical frameworks (Pigliucci 2017: 99).

Based on these assertions, Pigliucci argues that contemporary evolutionary biology is not undergoing a paradigm-shifting revolution, but rather a 'continuous expansion of both empirical knowledge and conceptual understanding' (Pigliucci 2017: 100), similar to those that, according to his view, would have taken place throughout the history of evolutionary biology since the Darwinian Revolution.

Pigliucci's analysis constitutes a valuable first assessment of the possible incommensurability between MS and EES. However, we believe that it is based on at least four dubious – or even downright incorrect – assumptions about incommensurability and the nature of the MS versus the EES controversy. More specifically:

(a) *Pigliucci seems to assume that incommensurability is a holistic phenomenon.*
(b) *Pigliucci's exposition of the debate between the MS and the EES overlooks the most contentious issues of the controversy.*
(c) *Pigliucci does not interpret observational incommensurability in the most meaningful way.*
(d) *Pigliucci assumes that incommensurability always implies a revolutionary process of paradigm shift.*

In the next section, we will address these four problems, showing how they compromise Pigliucci's main conclusion, namely that the MS and the EES are commensurable theoretical frameworks. In Sect. 4, we offer an alternative analysis

of incommensurability between the MS and the EES that seeks to overcome these problems in Pigliucci's analysis.

3 A Closer Look at the Problems of Pigliucci's Analysis

3.1 Incommensurability Is Not a Holistic Phenomenon

The first problematic assumption in Pigliucci's analysis is that he seems to assume that incommensurability is a global phenomenon, that is, a phenomenon that affects *all* methods, observations, and/or concepts of competing theoretical frameworks. Because Pigliucci finds at least some common methods, observations, and concepts between the MS and the EES, he concludes that both frameworks are not incommensurable. It is true that, in his chapter, Pigliucci asserts that if incommensurability is found in just one domain (either methodology, observations, or semantics), then 'a good argument can be made that a paradigm shift [i.e., incommensurability] is occurring' (Pigliucci 2017: 99). However, he seems to assume that *all* items in either of these classes have to be incommensurable in order for incommensurability to occur.[3]

We believe that this is a dubious interpretation of incommensurability. In our view, the fact that incommensurability can act as a proxy for radical departure does not imply that two incommensurable theoretical frameworks must be different in all of their methods, observations, and/or concepts, or that no points of continuity can be found between them. Even fully incommensurable paradigms share some items. This is so because, as Kuhn pointed out, incommensurability is not a holistic but a local phenomenon (Kuhn 2000a), meaning that it is generally restricted to a limited – yet especially important—set of items in each domain. Pigliucci seems to miss this point in his discussion of incommensurability, tacitly assuming that for incommensurability to occur there cannot be any overlap between the MS and the EES. A more thoughtfully Kuhnian interpretation must consider the local nature of incommensurability.

3.2 Analyses of Incommensurability Should Focus on the Most Controversial Aspects of the Dispute

The fact that incommensurability is a local phenomenon restricted to a limited – yet especially important – subset of methods, observations, and concepts force us to carefully reflect on which of these methods, observations, and concepts we should choose to perform our incommensurability analyses. Otherwise, the risk is to

[3] We thank an anonymous reviewer for pointing this to us.

conclude that two potentially incommensurable paradigms are commensurable just because we have focused on those methods, observations, or terms that both paradigms share. We believe that this is what happens in Pigliucci's analysis: When arguing against incommensurability, he focuses on the relatively uncontroversial parts of the MS versus EES debate, giving the impression that there is much more continuity between these two frameworks than there is. For example, two terms that according to Pigliucci show that the MS and the EES are semantically commensurable are 'species' and 'phenotype'. However, the fact that these terms are not under dispute does not mean that there is also agreement on the meaning of other crucial concepts.

Because of this, we believe that analyses of incommensurability should focus only on the most controversial items in a given scientific debate. Only by analysing the most controversial methods, observations, and concepts would we be able to determine whether there is incommensurability between the two theoretical frameworks involved. To detect what these items are, it is important to look not only at the conceptual structure of the theoretical frameworks involved as they are depicted in scientific publications but also at the way scientists from different paradigms debate in scientific meetings and confrontational papers. This is important because some methods, observations, or concepts that from a strictly theoretical point of view may not seem very problematic may turn out to be so when we observe how scientists debate about them. In fact, it could be argued that this is yet another problem with Pigliucci's analysis: he does not refer to the growing number of scientific papers in which proponents of the MS and the EES debate different points of their theories (Dickins and Rahman 2012; Mesoudi et al. 2013; Scott-Phillips et al. 2014; Laland et al. 2014; Gupta et al. 2017a, b; Feldman et al. 2017; Welch 2017; Laland 2017; Müller 2017; Futuyma 2017; Tanghe et al. 2018; Svensson 2023). These may give a better impression as to whether the MS and the EES are really incommensurable than a purely theoretical analysis. Although Kuhn did not specify where exactly to look for conducting incommensurability analyses, we believe that our proposal is fully in line with a Kuhnian view of science, since it places its emphasis on scientific practice rather than only on its purely formal elements (Rouse 2003).

3.3 Observational Incommensurability Is Really About Ontological Assumptions

The third problematic assumption in Pigliucci's analysis is related to his interpretation of observational incommensurability. Pigliucci seems to adhere to the most common interpretation of observational incommensurability, arguing that it means that 'what is considered a "fact" within one theoretical context may not be such in a different theoretical context' (Pigliucci 2017: 99). As we suggested in Sect. 2.1, there is no straightforward way to interpret observational incommensurability, given Kuhn's ambiguities on that matter, so Pigliucci cannot be blamed for interpreting

observational incommensurability in such a way. However, we believe that there is a more suggestive and Kuhnian way of interpreting observational incommensurability than that of Pigliucci.

According to this interpretation, observational incommensurability would amount to irreconcilable differences in very basic ontological assumptions; assumptions about which entities and processes populate the world. As Kuhn pointed out, each paradigm or theoretical framework includes a series of fundamental metaphysical beliefs that guide the interpretation of directly observable phenomena (i.e., 'models'). The ontological assumptions of competing paradigms depart significantly from each other to the point of being incompatible. For this reason, controversies regarding observational incommensurability do not revolve around 'what is considered a "fact"', as Pigliucci and many others suggest, but rather around how—that is, under which ontological assumptions—to interpret a *shared* set of facts. Due to its emphasis on ontological assumptions rather than empirical facts, we have elsewhere called this interpretation of observational incommensurability 'ontological incommensurability' (Gefaell and Saborido 2022).

This ontology-centred way of understanding observational incommensurability is supported by some examples discussed by Kuhn in *SSR*, such as that of the conceptions of space in Newtonian and relativistic physics. According to Kuhn, Newtonian scientists are 'embedded in a flat [space]', while Einsteinians inhabit '(. . .) a curved matrix of space' (Kuhn 1970: 150). The concept of space is not an observable phenomenon, but a very broad ontological category, and beliefs in a 'flat' or 'curved matrix' of space are not empirical claims, but metaphysical beliefs or assumptions that guide the interpretation of data collected by scientists. This example suggests that Kuhn understood observational incommensurability in a much more ontology-oriented way than traditional accounts of the concept assume. Recent attempts to develop observational incommensurability under a connectionist framework, such as that of Alexander Bird (2005, 2008), are also congruent with this interpretation of observational incommensurability as referring to the ontological assumptions of paradigms. We suggest that this way of interpreting observational incommensurability is not only closer to Kuhn's original position, but also may yield more interesting results when applied to scientific controversies.

3.4 Incommensurability Does Not Imply Paradigm-Shifting Revolutions

The fourth problematic assumption in Pigliucci's analysis is that he automatically links incommensurability with paradigm-shifting revolutions. However, a closer look at Kuhn's writings reveals that incommensurability does not necessarily imply a paradigm-shifting revolution, but it can also lead to other kinds of scientific change, particularly the formation of new scientific specialties (Kuhn 2000b, c; Politi

2019). This means that it is possible to argue in favour of incommensurability without committing to the inevitability of a paradigm shift.

Instead of a paradigm-shifting revolution, in the scientific specialisation scenario incommensurability leads to the splitting of a scientific community into two coexisting, yet completely autonomous, scientific communities. According to the most common interpretation, this process of scientific specialisation is normally triggered by anomalies (see Politi 2019). As Kuhn explains in *SSR*, at the time of their inception, all paradigms have anomalies or unsolved problems. Most of the time, these anomalies are overlooked by the majority of ordinary scientists, who instead focus on easier problems that do not call into question the established framework (what Kuhn called 'puzzles'), leaving anomalies for a time when the paradigm is ripe enough to resolve them. However, a minority of scientists still spend a great deal of time thinking and investigating these anomalies. This may lead to the development of new tools, concepts, or hypotheses, some of which may divert from the 'orthodox' point of view. If some of these tools, concepts, or hypotheses shed light onto some anomalies, then it may be the starting point of a specialisation process: they may further the divergence with established methods and concepts and may speed up the discovery of new phenomena that cannot be accounted for under the orthodox framework. On these occasions, the minority of scientists who initiated the above-mentioned changes may eventually end up becoming a separate scientific community, with its own tools and research agenda. When, after some time, this newly formed scientific community seeks to establish contact with its 'mother' community, differences in methods, assumptions, and concepts would prevent full communication; in other words, incommensurability would occur. In this sense, incommensurability would be the 'isolation mechanism' by which both communities lose contact with each other.[4]

Given that incommensurability can be associated with different kinds of scientific change, if one obtains evidence in favour of incommensurability, it becomes necessary to determine independently what mode of scientific change (paradigm shift, scientific specialisation, or other) accompanies incommensurability in that particular case. Pigliucci is wrong in automatically assuming that incommensurability necessarily implies a paradigm-shifting revolution.

[4]The reference to biological vocabulary, such as 'isolation mechanism', is no coincidence, since Kuhn employed the 'speciation' metaphor to understand scientific specialisation. As in allopatric speciation, where one species eventually becomes two separate ones after it is split up in two populations by a geographical barrier, making selective pressures to act differently in each population, scientific specialisation would also involve an isolation mechanism (incommensurability) and a divergence in selective pressures (different problems tackled by each group of scientists).

4 An Alternative Analysis of Incommensurability Between the MS and the EES

Due to the aforementioned problems in his account, Pigliucci asserts that there is no incommensurability between the MS and the EES at the methodological, observational, and semantic levels. Do we get the same conclusion when we try to avoid these problems? We argue that this is not the case: we believe that when these problems are addressed, evidence suggests that the MS and the EES are incommensurable theoretical frameworks. To show why, let us follow the same approach as Pigliucci, considering methods, observations, and concepts separately.[5]

4.1 Methodological Incommensurability Between the MS and the EES

We have already seen that Pigliucci defends that there is no methodological incommensurability between the MS and the EES. We disagree, as we detect at least four instances of methodological incommensurability between these two theoretical frameworks.

The first one has to do with their research agenda, or, in other words, the kinds of problems that each theoretical framework considers of high priority and most scientific interest. The MS research agenda is markedly influenced by population genetics. MS evolutionary biologists are concerned with, for instance, problems related to effective census (*Ne*) estimation, the modeling of genetic drift on allele frequencies, or the effects of consanguinity on the average fitness of populations, to name a few examples. In this sense, we believe most adherents to the MS still agree with Dobzhansky's claim that '(. . .) the mechanisms of evolution constitute problems of population genetics' (Dobzhansky 1937: 11–12).[6]

On the contrary, supporters of the EES do not grant population genetics such importance in their research agenda. Instead, they tend to focus on developmental processes and their role in evolution, as well as what can be called 'agentic' problems, that is, those that have to do with intentional or agential behaviours performed by organisms and their impact on evolution (such as, for example,

[5] We will also follow Pigliucci (2017) in understanding the EES as the framework exposed in Laland et al. (2015), as we agree with him that this paper 'is both more focused and more systematic [in its depiction of the EES] than previous attempts' (Pigliucci 2017: 96). As for the MS, we understand it as the theoretical framework depicted in mainstream evolutionary biology textbooks, such as Ridley (2004), Freeman and Herron (2004), or Futuyma (2009). This move is not a coincidence, since Kuhn stressed the role of textbooks in establishing dominant paradigms (Kuhn 1963).

[6] Other more recent vindications of the importance of population genetics for the MS can be seen in Ridley (2004: 93), Freeman and Herron (2004: 141), Lynch (2007), or Futuyma (2009: 220).

many actions and processes that fall under the label of 'niche construction'; see Odling-Smee et al. 2003; Laland et al. 2015). The different research agendas of each framework also influence the kinds of organisms that are used to conduct research in the MS and the EES. In this sense, the EES is much more open to novel model organisms that are better suited to the study of developmental processes, such as *Onthophagus* beetles (Baedke et al. 2020), whereas advocates of the MS rely more on traditional model organisms, such as *Drosophila* flies or guppy fishes. Of course, these differences in their research agendas and model organisms are far from absolute. We are not arguing that supporters of the EES completely disregard population genetics or that the MS gets no interest at all in developmental processes, but that there is a major shift of emphasis in their research priorities. We believe that this shift is of much more relevance than Pigliucci thinks and that it is responsible for many conflicts between proponents of the MS and the EES.

The second instance of methodological incommensurability has to do with what we have called *explanatory preferences* (Gefaell and Saborido 2022). Explanatory preferences can be defined as the kinds of entities and processes that a given theoretical framework prioritises as *explanantia* in scientific explanations. In the life and mind sciences, there are two widely known explanatory preferences: explanatory externalism and explanatory internalism (Godfrey-Smith 1996). Explanatory externalism refers to 'explanations of properties of organic systems in terms of properties of their environments', while explanatory internalism refers to '[e]xplanations of one set of organic properties [in this case, the capacity for evolution] in terms of other internal or intrinsic properties of the organic system' (Godfrey-Smith 1996: 30).

Following these definitions, we believe that the MS leans toward explanatory externalism, while the EES toward explanatory internalism (e.g., Fábregas-Tejeda and Vergara-Silva 2018). This difference in explanatory styles can be seen, for instance, in the way each theoretical framework explains adaptations, that is, the fit between organisms and their environments: While the MS usually conceives these as the direct result of selective pressures imposed on organisms by – external – ecological factors, the EES emphasises the role of developmental plasticity and niche construction along natural selection in achieving such a fit (Scott-Phillips et al. 2014; Laland et al. 2015). This move from explanatory externalism to explanatory internalism is acknowledged by proponents of the EES themselves:

> '(. . .) [the EES] moves the focus of evolutionary explanation from the external and contingent to the internal and inherent. It posits that the causal basis for phenotypic form resides not in population dynamics or, for that matter, in molecular evolution, but instead in the inherent properties of evolving developmental systems' (Müller 2007: 947–948).

The third instance of methodological incommensurability between the MS and the EES involves epistemic values. Kuhn argued that although all natural scientists, *qua* scientists, share a series of epistemic values that are constitutive of science itself, their ranking, or the relative importance granted to each of them, varies between theoretical frameworks (Kuhn 1977). This observation also holds true for the MS and the EES. For instance, the relative importance given to the epistemic value of

simplicity is different in the MS and the EES: supporters of the MS tend to give this value more importance than advocates of EES. An example of this can be seen in Douglas Futuyma's claim that 'simpler explanations are generally preferred over more complex (and vague) hypotheses, unless these are supported by evidence' (Futuyma 2017: 8), referring to EES-styled explanations. Compare this statement with advocates of the EES claiming that explanations of the MS are often 'too simple' or 'impoverished', in a clear derogation of the value of simplicity (Scott-Phillips et al. 2014: 1233–1240). Rather than promoting simplicity, advocates of the EES grant less idealised explanations a higher epistemic value—even if these explanations are much more difficult to model because they involve many more causal factors than those of the MS.[7]

Finally, the fourth instance of methodological incommensurability has to do with the different views each theoretical framework has about the aims and general outlook of evolutionary biology as a science. In this regard, advocates of the EES have a much more favourable opinion toward scientific pluralism than their MS counterparts. Several quotes illustrate this; for example: 'We believe that a plurality of perspectives in science encourages the development of alternative hypotheses, and stimulates empirical work' (Laland et al. 2014: 164); 'We believe that a plurality of perspectives in science is healthy (. . .)' (Laland et al. 2015: 10; see also Laland 2018; Uller et al. 2019). On the contrary, adherents to the MS, with its insistence that a new evolutionary theory is not necessary (even when supporters of the EES explicitly say that they do not want to supersede the MS, but only to turn evolutionary biology into a plural discipline), seem to be much more inclined toward scientific monism or the unity of science. Theoretical unification was a major epistemic goal during the inception of the MS (Smocovitis 1992), and it still probably is for many orthodox evolutionary biologists (but see Svensson 2023).

4.2 Observational (That Is, Ontological) Incommensurability Between the MS and the EES

Although the observational basis of the MS and the EES is mostly shared, the ontological assumptions underlying their interpretation of phenomena are significantly different and in many cases difficult to reconcile. We detect at least two instances of ontological incommensurability: one related to the nature of organisms and the other to the nature of causality. Regarding the nature of organisms, the MS and the EES assume two significantly different metaphysical beliefs: On the one hand, the ontology of the MS is gene-centred, reflecting the population genetic basis

[7] This can be seen, for instance, in Uller et al. (2019). These authors contrast evolutionary explanations based on natural selection acting on genes with evolutionary explanations based on developmental plasticity, physiology, and behaviour. The latter explanations include many more causal factors, and assume more complex causal chains, than the former.

of the MS (Provine 1971). The MS conceives organisms as gene-carriers, or, as Richard Dawkins has put it, 'vehicles' for genes (Dawkins 1976; see also Hull 1980). Genes are the entities that ultimately evolve, not organisms.[8] On the other hand, for the EES, organisms occupy the centre stage. Organisms are not reducible to genes and instead act as integrated wholes. They share a set of special properties that, although not incompatible with physical laws, transcend them. They have the capacity for agency and autonomy and, for at least some EES-friendly authors, cannot be studied under a purely mechanical framework (e.g., Walsh 2015). The organicism espoused by the EES has several theoretical and methodological consequences. Particularly, it can be argued that the fact that adherents of the EES confer the study of developmental biology and agentic behaviour such importance stems from their metaphysical assumption that organisms are integrated wholes with special properties which are not reducible to mechanical ones.

The second instance of ontological incommensurability between the MS and the EES is the nature of causality (e.g., Laland et al. 2011, 2013; Martínez and Esposito 2014; Fábregas-Tejeda and Vergara-Silva 2018; Uller and Helanterä 2019). A core ontological assumption of the MS is the so-called proximate versus ultimate dichotomy of causation, originally put forward by one of the MS architects, Ernst Mayr, in the 1960s (Mayr 1961). This distinction implies that the proximate causes of biological traits (their developmental and physiological basis) have nothing to do with their evolutionary causes. Conversely, followers of the EES explicitly reject the proximate versus ultimate dichotomy. In fact, it can be argued that their entire research project is based on the premise that this dichotomy does not hold true and that what under such a framework are considered proximate mechanisms can indeed have a lasting influence on evolution (Laland et al. 2011, 2013, 2015).

Many misunderstandings that surround the debate around the convenience of the EES involve the status of the proximate versus ultimate dichotomy. When proponents of the EES vindicate the importance of development and behaviour for evolution, supporters of the MS reply that these are not relevant processes for a theory of evolution because they constitute proximate causes (e.g., Dickins and Rahman 2012; Futuyma 2017). The EES camp rejects this charge because they deny the validity of such a dichotomy (e.g., Laland et al. 2013, 2015). This generates fierce disagreements and communication failures between both camps.

In addition to the dispute over the status of the proximate versus ultimate dichotomy, other ontological assumptions regarding the nature of causation also generate disagreement between the MS and the EES. These include multilevel causation (Noble 2012; Martínez and Esposito 2014), and the so-called 'fragmentation of evolution', the notion tacitly assumed by supporters of the MS according to which the processes involved in inheritance, variation, and differential fitness are

[8] It could be replied that the gene-centred ontology was disputed by some of the early proponents of the MS, notably Ernst Mayr (1963). However, even authors such as Mayr can be seen as committed to a gene-centred view of the organism. This is so because although Mayr has been known for his criticism of 'bean-bag genetics', he has also argued that organisms are the result of the unfolding of a genetic program (Mayr 2004).

causally independent. Some proponents of the EES reject this fragmentation and claim that all of these processes are causally intertwined (Walsh 2015, 2016; Uller and Helanterä 2019).

4.3 Semantic Incommensurability Between the MS and the EES

Although we agree with Pigliucci that the meaning of many terms is shared by the MS and the EES, we believe that other concepts are more problematic and cast doubt on Pigliucci's conclusion. For example, a potential example of semantic incommensurability between these theoretical frameworks is the very concept of 'evolution'. Under the MS, evolution is understood as changes in gene or allele frequencies (e.g., Dobzhansky 1937; Freeman and Herron 2004; Futuyma 2009). The EES, for its part, usually employs (supposedly) much broader definitions of evolution, such as, 'transgenerational change in the distribution of heritable traits of a population' (e.g., Laland et al. 2015: 2). These differences in the way each theoretical framework defines a concept as important as 'evolution' sometimes gives rise to communication failures. For example, for many supporters of the MS, niche construction cannot count as an evolutionary mechanism (as advocates of the EES believe) because it does not alter gene frequencies (Scott-Phillips et al. 2014). On the contrary, given that the EES defines evolution in an allegedly much broader way, including changes in ecological inheritance brought up by niche construction, its advocates have no problem in considering this process as an evolutionary mechanism.[9]

Some may object that what we have just argued is perfectly compatible with Pigliucci's account, as the examples discussed above suggest that evolution under the EES is understood in a *much broader way than* – yet *not incommensurable with* – the corresponding concept in the MS. To this objection, our reply is that, despite appearances, the meaning of evolution in the EES is not a mere extension of that concept in the MS. The reason why this is so is that each conception of evolution carries with it significantly different metaphysical assumptions (particularly, those discussed in the previous section). More specifically, 'evolution' in the MS assumes the causal independence of inheritance, variation, and differential fitness (the 'fragmentation of evolution'; Walsh 2015, 2016; Uller and Helanterä 2019), while the EES stresses their interactions. For this reason, a theoretical integration between the notions of 'evolution' of the MS and the EES is difficult and would require a great deal of conceptual work.

In sum, contrary to Pigliucci, we believe that there are several cases of methodological, observational (i.e., ontological), and semantic incommensurability between

[9] Similar deep disagreements can be found with respect to other important evolutionary concepts, such as, for instance, 'inheritance' (Danchin et al. 2011; Danchin and Pocheville 2014; Uller and Helanterä 2017; see Gefaell and Saborido 2022).

Table 1 A summary of cases of incommensurability between the MS and the EES

Incommensurability between the MS and the EES	
Methodological	Different research agenda
	Different explanatory preferences
	Different ranking of epistemic values
	Different views about science itself
Observational (i.e., ontological)	Different views on organisms
	Different views on causality
Semantic	Different understanding of evolution and other central concepts (particularly inheritance)

the MS and the EES. As for the methods, these frameworks differ in their research agenda, explanatory preferences, epistemic values, and view of science. With respect to ontology, the MS and the EES have significantly different views on the nature of organisms and causality. Finally, both frameworks espouse different notions of evolution (see Table 1).

5 What Type of Scientific Change Is Currently Going on?

So far, our analysis has left the question of which mode of scientific change accompanies incommensurability unanswered. In Sect. 3.4, we argued that from a Kuhnian point of view, incommensurability can prompt either a paradigm-shifting revolution or a process of scientific specialisation. So, which of the two is more likely to occur as a consequence of the controversy between the MS and the EES?

Although we believe that it is too early to rule out any possibility, there are at least two reasons not to commit to a paradigm-shifting scenario. The first reason has to do with the reluctance of most evolutionary biologists to join the EES. We do not have statistical data on the percentage of evolutionary biologists sympathetic to the EES, but our guess is that they constitute a small minority. Many evolutionary biologists working under the MS do not even know about the existence of an alternative theoretical framework, the EES being much better known and far more widely discussed in philosophical circles than in purely scientific ones. Unless the proponents of the EES present spectacular new achievements and run large outreach campaigns, it is not likely that this situation of relative ignorance by most practicing evolutionary biologists will change in the short term. It seems that advocates of the EES have given up the task of pushing for its own agenda. Although Kuhn (1970) argued that groups affected by a paradigm shift often number around one hundred or fewer members, there is still a long way to 'convert' most evolutionary biologists to the EES. For a paradigm-shifting revolution to occur, most scientists of the old paradigm must eventually abandon it and join the new paradigm. And given current circumstances, this is unlikely to happen in the near future.

The second reason we have for doubting a scientific revolution is the very intentions of the supporters of the EES themselves: most of them have explicitly said that they do not want to replace the MS, but only to turn evolutionary biology into a plural science.[10] As we have seen in Sect. 4.1, most supporters of the EES are committed scientific pluralists; they believe in the value of the co-existence of different perspectives, meaning that they only want to have their space within evolutionary biology alongside that of the MS to conduct their own research agenda, with their very own methods, concepts, and assumptions. This is even acknowledged by some critics of the EES (Svensson 2023). At least for now, we have no compelling reasons to think that these are not their real intentions. Therefore, if the supposedly 'revolutionary' scientists do not want to break with the MS, then the scenario of a paradigm shift weakens further.

But now we may ask: Does the above necessarily imply that scientific specialisation is going to take place? Although it might certainly seem more plausible at first sight and although it is compatible with the pluralist leanings of advocates of the EES, it is nonetheless too soon to conclude that the EES will turn into a separate specialty within the realm of evolutionary biology, perhaps engulfing evo-devo and other related subdisciplines. Besides, there is at least one potential caveat against this possibility, namely that the MS and the EES, far from being completely autonomous from each other in terms of their objects of study, in some respects compete for the explanation of the same evolutionary events. This can be seen in the case of niche construction, where the EES and the MS compete for the explanation of phenomena such as, for instance, the evolution of lactose tolerance in humans (Scott-Phillips et al. 2014). If two theoretical frameworks compete for the explanation of the same phenomena, then it is highly implausible that they will end up being completely autonomous scientific specialties. Points of friction and controversy will likely emerge.[11] This obliges us to remain agnostic about which kind of scientific change may follow from the incommensurability between the MS and the EES.

6 Conclusion

In this chapter, we have critically analysed Pigliucci's recent proposal to apply the Kuhnian notion of incommensurability to the MS versus EES controversy. We have argued that, while valuable, his analysis is based on four problematic assumptions

[10] Only a few advocates of the EES or EES-friendly authors adopt a more radical stance, arguing in favour of the substitution of the MS by an EES or equivalent (e.g., Noble 2015).

[11] A third plausible scenario that has been put forward recently is to conceive the EES as a Kuhnian reformulation of the MS (Tanghe et al. 2021: 15). This entails a sort of middle-ground between a paradigm-shifting revolution and a business-as-usual extension of the established paradigm. Viewing the EES as a Kuhnian reformulation of the MS is compatible with a local form of incommensurability like the one we have suggested in this chapter. However, what a Kuhnian reformulation entails and how to precisely detect it in particular contexts is something not yet clear.

that call into question his main conclusion, i.e., that there is no incommensurability between the MS and the EES. After showing what these problematic assumptions are, we have outlined our own analysis of incommensurability, concluding that there are several instances of incommensurability between the EES and the MS. However, we have also argued that our analysis of incommensurability does not imply that a paradigm shift is currently taking place in evolutionary biology.

If we are correct in our analysis, the existence of incommensurability between the MS and the EES leaves the door open for a number of additional philosophical problems that demand resolution. These include the nature and implications of deep disagreements in science, the rationality of theory evaluation, or the question of scientific realism (Psillos 2018). Elsewhere we have outlined potential avenues for the resolution of at least some of these problems (Gefaell and Saborido 2022), but a full exploration of them should be a matter of future investigation for philosophers and philosophically-inclined evolutionary biologists.

Acknowledgments We thank Mariano Sanjuán and José Manuel Viejo for the opportunity to participate in this volume. We also thank two anonymous reviewers for their constructive criticism on the early versions of this chapter, which helped to significantly improve it. Juan Gefaell is funded by a Xunta de Galicia Predoctoral Research Contract (ED481A-2021/274). Cristian Saborido is grateful for funding from the Spanish Ministry of Science (PID2021-128835NB-100 research project).

References

Andersen H, Hepburn B (2013) Scientific change. In: The internet encyclopedia of philosophy. https://iep.utm.edu/scientific-change/. Accessed 30 May 2022

Baedke J, Fábregas-Tejeda A, Vergara-Silva F (2020) Does the extended evolutionary synthesis entail extended explanatory power? Biol Philos 35:20

Bird A (2005) Naturalizing Kuhn. Proc Aristot Soc 105:109–127

Bird A (2008) Incommensurability naturalized. In: Soler L, Sankey H, Hoyningen-Huene P (eds) Rethinking scientific change and theory comparison. Springer, Dordrecht, pp 21–39

Chang H (2013) Incommensurability: revisiting the chemical revolution. In: Kindi V, Arabatzis T (eds) Kuhn's structure of scientific revolutions revisited. Routledge, pp 153–178

Danchin E, Charmantier A, Champagne FA, Mesoudi A, Pujol B, Blanchet S (2011) Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. Nat Rev Genet 12:475–486

Danchin E, Pocheville A (2014) Inheritance is where physiology meets evolution. J Physiol 592: 2307–2317

Dawkins R (1976) The selfish gene. Oxford University Press, New York

Dickins TE, Rahman Q (2012) The extended evolutionary synthesis and the role of soft inheritance in evolution. Proc R Soc B Biol Sci 279(1740):2913–2921

Dobzhansky T (1937) Genetics and the origin of species. Columbia University Press, New York

Fábregas-Tejeda A, Vergara-Silva F (2018) Hierarchy theory of evolution and the extended evolutionary synthesis: some epistemic bridges, some conceptual rifts. Evol Biol 45:127–139

Feldman MW, Odling-Smee J, Laland KN (2017) Why Gupta et al.'s critique of niche construction theory is off target. J Genet 96:505–508

Feyerabend PK (1962) Explanation, reduction and empiricism. In: Feigl H, Maxwell G (eds) Scientific explanation, space and time. Minnesota studies in the philosophy of science, vol 3. University of Minnesota Press, Minnesota, pp 28–97
Freeman S, Herron JC (2004) Evolutionary analysis. Pearson, New Jersey
Futuyma DJ (2009) Evolution. Sinauer Associates, Sunderland
Futuyma DJ (2017) Evolutionary biology today and the call for an extended synthesis. Interface Focus 7:20160145
Gefaell J, Saborido C (2022) Incommensurability and the extended evolutionary synthesis: taking Kuhn seriously. Eur J Philos Sci 12:24
Godfrey-Smith P (1996) Complexity and the function of mind in nature. Cambridge University Press, Cambridge
Gupta M, Prasad NG, Dey S, Joshi A, Vidya TNC (2017a) Niche construction in evolutionary theory: the construction of an academic niche? J Genet 96:491–504
Gupta M, Prasad NG, Dey S, Joshi A, Vidya TNC (2017b) Feldman et al. do protest too much, we think. J Genet 96:509–511
Hoyningen-Huene P (1990) Kuhn's conception of incommensurability. Stud Hist Philos Sci Part A 21:481–492
Hoyningen-Huene P (1993) Reconstructing scientific revolutions. Thomas Kuhn's philosophy of science. The University of Chicago Press, Chicago
Hoyningen-Huene P, Sankey H (eds) (2001) Incommensurability and related matters. Springer
Hull DL (1980) Individuality and selection. Annu Rev Ecol Syst 11:311–332
Kuhn TS (1963) The function of dogma in scientific research. In: Crombie A (ed) Scientific change. Heineman Educational Books, London, pp 347–369
Kuhn TS (1970) The structure of scientific revolutions. The University of Chicago Press, Chicago
Kuhn TS (1977) The essential tension: selected studies in scientific tradition and change. The University of Chicago Press, Chicago
Kuhn TS (2000a) Commensurability, comparability, communicability. In: Conant J, Haugeland J (eds) The road since *the structure*. The University of Chicago Press, Chicago, pp 33–57
Kuhn TS (2000b) The trouble with the historical philosophy of science. In: Conant J, Haugeland J (eds) The road since *the structure*. The University of Chicago Press, Chicago, pp 105–120
Kuhn TS (2000c) Afterwards. In: Conant J, Haugeland J (eds) The road since *the structure*. The University of Chicago Press, Chicago, pp 224–252
Laland KN (2017) Schism and synthesis at the Royal Society. Trends Ecol Evol 32:316–317
Laland KN (2018) Evolution unleashed. Aeon. https://aeon.co/essays/science-in-flux-is-a-revolution-brewing-in-evolutionary-theory. Accessed 26 Oct 2021
Laland KN, Odling-Smee J, Hoppitt W, Uller T (2013) More on how and why: cause and effect in biology revisited. Biol Philos 28:719–745
Laland KN, Sterelny K, Odling-Smee J, Hoppitt W, Uller T (2011) Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? Science 334:1512–1516
Laland KN, Uller T, Feldman MW et al (2014) Does evolutionary theory need a rethink? Nature 514:161–164
Laland KN, Uller T, Feldman MW et al (2015) The extended evolutionary synthesis: its structure, assumptions and predictions. Proc R Soc B Biol Sci 282:20151019
Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci USA 104:8597–8604
Martínez M, Esposito M (2014) Multilevel causation and the extended synthesis. Biol Theory 9: 209–220
Mayr E (1963) Animal species and evolution. Harvard University Press, Harvard
Mayr E (1961) Cause and effect in biology. Science 134:1501–1506
Mayr E (1994) The advance of science and scientific revolutions. J Hist Behav Sci 30:328–334
Mayr E (2004) What makes biology unique? Cambridge University Press, Cambridge
Mesoudi A, Blanchet S, Charmantier A et al (2013) Is non-genetic inheritance just a proximate mechanism? A corroboration of the extended evolutionary synthesis. Biol Theory 7:189–195

Müller GB (2007) Evo-devo: extending the evolutionary synthesis. Nat Rev Genet 8:943–949
Müller GB (2017) Why an extended evolutionary synthesis is necessary. Interface Focus 7: 20170015
Noble D (2012) A theory of biological relativity: no privileged level of causation. Interface Focus 2: 55–64
Noble D (2015) Evolution beyond neo-Darwinism: a new conceptual framework. J Exp Biol 218:7–13
Odling-Smee J, Laland KN, Feldman MW (2003) Niche construction: the neglected process in evolution. Princeton University Press, Princeton
Pigliucci M (2007) Do we need an extended evolutionary synthesis? Evolution 61:2743–2749
Pigliucci M (2009) An extended synthesis for evolutionary biology. Ann N Y Acad Sci 1168:218–228
Pigliucci M (2012) Biology's last paradigm shift. Paradigmi 3:45–58
Pigliucci M (2017) Darwinism after the modern synthesis. In: Delisle RG (ed) The Darwinian tradition in context: research programs in evolutionary biology. Springer International Publishing, pp 89–103
Pigliucci M, Müller GB (eds) (2010) Evolution, the extended synthesis. The MIT Press, Cambridge
Politi V (2018) Scientific revolutions, specialization and the discovery of the structure of DNA: toward a new picture of the development of the sciences. Synthese 195:2267–2293
Politi V (2019) Specialisation and the incommensurability among scientific specialties. J Gen Philos Sci 50:129–144
Provine WB (1971) The origins of theoretical population genetics. The University of Chicago Press, Chicago
Psillos S (2018) Realism and theory change in science. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/realism-theory-change/. Accessed 31 May 2022
Ridley M (2004) Evolution. Wiley-Blackwell, Oxford
Rouse J (2003) Kuhn's philosophy of scientific practice. In: Nickles T (ed) Thomas Kuhn. Cambridge University Press, Cambridge, pp 101–121
Sankey H (1993) Kuhn's changing concept of incommensurability. Br J Philos Sci 44:759–774
Scott-Phillips TC, Laland KN, Shuker DM et al (2014) The niche construction perspective: a critical appraisal. Evolution 68:1231–1243
Smocovitis VB (1992) Unifying biology: the evolutionary synthesis and evolutionary biology. J Hist Biol 25:1–65
Svensson EI (2023) The structure of evolutionary theory: beyond neo-Darwinism, neo-Lamarckism and biased historical narratives about the modern synthesis. In: Dickins TE, Dickins BJA (eds) Evolutionary biology: contemporary and historical reflections upon core theory. Springer, pp 173–217
Tanghe KB, De Tiège A, Pauwels L, Blancke S, Braeckman J (2018) What's wrong with the modern evolutionary synthesis? A critical reply to Welch (2017). Biol Philos 33:23
Tanghe KB, Pauwels L, De Tiège A, Braeckman J (2021) Interpreting the history of evolutionary biology through a Kuhnian prism: sense or nonsense? Perspect Sci 29:1–35
Uller T, Feiner N, Radersma R et al (2019) Developmental plasticity and evolutionary explanations. Evol Dev 2:47–55
Uller T, Helanterä H (2017) Heredity and evolutionary theory. In: Walsh D, Huneman P (eds) Challenging the modern synthesis: adaptation, development, and inheritance. Oxford University Press, Oxford, pp 280–316
Uller T, Helanterä H (2019) Niche construction and conceptual change in evolutionary biology. Br J Philos Sci 70:351–375
Walsh DM (2015) Organisms, agency, and evolution. Cambridge University Press, Cambridge
Walsh DM (2016) Challenges to evolutionary theory. In: Humphreys P (ed) The Oxford handbook of philosophy of science. Oxford University Press, Oxford, pp 671–694
Welch JJ (2017) What's wrong with evolutionary biology? Biol Philos 32:263–279
Wray KB (2005) Rethinking scientific specialization. Soc Stud Sci 35:151–164
Wray KB (2011) Kuhn's evolutionary social epistemology. Cambridge University Press, Cambridge

Ontologies in Evolutionary Biology: The Role of the Organism in the Two Syntheses

David Cortés-García and Arantza Etxeberria Agiriano

Abstract This paper examines evolutionary ontologies from Darwin's work to the genesis and maturation of the Modern Evolutionary Synthesis, followed by the onset of the more inclusive framework of the Extended Evolutionary Synthesis. We show how, in an attempt to unify different biological fields under evolutionary principles, the first synthetic theory of evolution progressively disregarded the relevance of organismic-level properties and processes. Yet, failure to reduce the systemic nature and ecological dynamics of the organism (including properties of agency and organization) to that framework raises some important drawbacks. In particular, the two fundamental dimensions of developmental and ecological dynamics were largely neglected. These are the ones that highlight the relational properties of organisms and lead to recent views in, respectively, Evo-Devo and Niche Construction Theory. On this account, we argue that these two aspects illuminate how, while the Modern Synthesis became increasingly reductionist and monistic, the Extended Synthesis is currently being constituted by a pluralistic array of models capable of accommodating different ontological levels, among which that of organisms stands out due to its flexibility and potential inclusiveness.

Keywords Organizational and evolutionary principles · Darwin · Modern Synthesis · Evo-Devo · Niche Construction Theory · Extended Synthesis · Theory Integration · Agency · Organicism

D. Cortés-García (✉) · A. Etxeberria Agiriano
Department of Philosophy, University of the Basque Country, Donostia – San Sebastián, Spain

IAS Research Group for Life, Mind and Society, University of the Basque Country, Donostia – San Sebastián, Spain
e-mail: david.cortes@ehu.eus; arantza.etxeberria@ehu.eus

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_9

1 Introduction

Ontologies are crucial aspects of scientific theories, as they greatly influence further research and scientific practices. In the particular case of evolutionary biology, there have existed different understandings of the nature of evolving entities, as well as of the causal processes they undergo.

One of these ontological points at issue questions the kind of change that constitutes evolution: Does it happen constantly, in small, incremental steps, or suddenly, in moments of catastrophe? Is it directional? Which entities could/do evolve? Is evolution peculiar to life and of changes occurring in organic systems, or is it grounded in more basic properties of matter? Does it entail an increase in complexity? Some of the recent debates in evolutionary biology can be seen as discussions about the kind of change we call evolution and about the nature of the entities that undergo evolutionary change. Just as most complicated discussions, these questions rely on a metaphysical ground: as Stephen J. Gould pointed out[1] (Gould 1982: 383), the issue is whether the living world is constantly changing, and therefore structure is a mere incarnation that arises at particular times, or rather structure is a primary, constraining event for which change is difficult and, if it occurs, it occurs rapidly. Derivationally, the confrontation is also epistemological and, in fact, has long roots in the history of controversies in defense of different research programs in biology.

A related but not fully identical debate concerns the organized nature of the entities that evolve, in contrast with designed features sometimes associated to them. An important aspect for today's biological theory is the role of the organism[2] in evolution: although in earlier stages of evolutionary biology evolution was straightforwardly attributed to organisms, later on theoretical accounts referred to smaller and more fundamental units, such as genes, which became the main entities. As a consequence of this, the role of the organism was progressively lost in evolutionary biology, but recently many authors have maintained that it needs to be recovered (Baedke 2019; Etxeberria and Umerez 2006; Nicholson 2014; Ruiz-Mirazo et al. 2000).

The explanatory frameworks of the theory of evolution, especially in what concerns the two syntheses – the so-called *Modern* Synthesis and *Extended* Synthesis of evolution – their scientific endeavors, postulates, and assumptions about the living world, have lately been the target of strong discussions, including demands for

[1] "In the largest sense, this debate [referring to gradualism-punctuationism] is but one small aspect of a broader discussion about the nature of change: is our world (to construct a ridiculously oversimplified dichotomy) primarily one of constant change (with structure as a mere incarnation of the moment), or is structure primary and constraining, with change as a 'difficult' phenomenon, usually accomplished rapidly when a stable structure is stressed beyond its buffering capacity to resist and absorb" (Gould 1982: 383).

[2] The notion of "organism" is a neologism appearing at the end of the 1690s to emphasize the *organized* nature of some entities (not only living beings), as opposed to mechanisms (see, e.g., Cheung 2010).

expansions in different dimensions (or for the substitution) of the Modern Evolutionary Synthesis (MES) by a more comprehensive framework: purportedly the Extended Evolutionary Synthesis (EES) (Depew and Weber 2011; Fábregas-Tejeda and Vergara-Silva 2018a, b; Müller 2017; Pigliucci 2007, 2009).

In this paper, we propose that the main discrepancies between the Modern Synthesis and the Extended Synthesis need to be examined in the light of the ontologies they comprise, in particular with respect to the conceptualization of the organism and the role which is assigned to it in evolution. First, we consider the view that the framework within which Darwinian theory developed was organism-centered and examine how that perspective started to shift as inheritance was reconceived (Sect. 2). We then recall that early efforts to elaborate a unified theoretical biology that would provide the foundations of the field were carried out in organizational terms and only later were reconfigured in evolutionary terms, which gave the Modern Synthesis its particular characteristics (Sect. 3), including the main ontological problems arising, respectively, from structural and developmental criticisms of radical gradualism and genetic reductionism (Sect. 4). As for the ontological commitments of the Modern Synthesis during its maturation, it is noticeable that the organism was increasingly pushed to a marginal position with no causal relevance or explanatory role, as organismic biology was being progressively replaced by a gene-centered view of life. This outcome was challenged by developmental and ecological research and theories which contribute to an organism-centered revision of evolutionary biology (Sect. 5), so that questions which are being explored by the Extended Synthesis more recently can be conceived as a "return of the organism" to the current theory of evolution (Sect. 6). Finally, we reach a conclusion regarding the discordances between the two syntheses from an ontological standpoint focused on organisms.

2 Darwinian Gradualism, Inheritance, and the Organism

The publication of Charles Darwin's *On the Origin of Species* in 1859 was pivotal in driving modern evolutionary thinking. Darwin's theory of evolution is usually understood to be built upon two major pillars, which are logically independent from each other: the idea of *common ancestry* of all species, according to which all living forms and taxa are genealogically related to one another, and *natural selection* as the causal process involved in the adaptation of organisms to their environments (Sober and Orzack 2003). The former addresses the study of how forms and morphologies are mutually dependent, conforming historical contingent trajectories, while the latter seeks to understand the functions of organic parts and endowments as strategies to increase fitness values (i.e., differential survival or reproduction) in the environment. On both of these fronts Darwin assumes a *gradualist* and *organismic* view of the relevant processes.

Gradualism is a fundamental component of Darwin's evolutionary ontology, which conceives of nature as a continuum of organismal forms, and of evolution

as resulting from the accumulation of small changes. The idea that evolution proceeds through slow and gradual selection was consistent with his field observations and his experimental work on domestic races. Accordingly, he assumed that intermediate stages must have existed between any two species that are distinguishable as such. Hence, species form a continuum in which there are no notable differences between the category of species and mere variants or races. However, Darwin's gradualism was, in fact, challenged by the demands of addressing heredity, and, in turn, the way in which the question of heredity was resolved brought the organismic ontology to an end.

On the other hand, the organism played a prominent role in Darwin's ontology: due to their self-organizing character, the very nature of organisms is just as relevant as (or more than) the nature of external conditions. In this regard, it has been noted that Darwin's understanding of adaptation differs from Lamarck's: whereas the latter conceives of it as an immediate and directed response by organisms to their surroundings, Darwin would contend that, apart from being sensitive to those conditions, organisms are autonomous and self-regulating, thus somehow creating the necessary conditions for selection (Brooks 2000). This aspect was later blurred, "as biologists focused more attention on parts of organisms, and less on organisms as wholes" (Brooks 2000: 259). Darwin's consideration of organisms also entails that they engage materially with their environment and constructively produce the conditions for selection. Thus, a clear sense of the organism can be appreciated in his writings: Darwin understands evolution as the alteration of the features and possibilities of organisms and their surroundings over time. He conveys this systemic view of the nature of organisms neatly throughout his work.[3] Denis Walsh has recently elaborated ideas along these lines which imply that, for Darwin, adaptation is related to the "purposive activities of organisms" (Walsh 2015: 159).

The new understanding of heredity that arose after Darwin' theory of evolution triggered a cascade of changes in attitudes both in the scientific and in the public understanding of life that ultimately led to the departure from the organism-based ontology. At the time, explanations of inheritance in sexual reproduction considered that the characters of the parents blended in the offspring, something that has been considered to be incompatible with the Darwinian theory of evolution and in need of a better account. Indeed, blending inheritance would eliminate the accumulation of variation, thus making natural selection ineffective. Besides, this theory is unable to explain how characters may remain latent in a lineage for one or more generations and then reappear expressed again (Bowler 2003; Meloni 2016). On his part, Darwin developed his own "theory of pangenesis" according to which cells of all parts of the body produce their own minute gemmules which may transmit their features when brought together in the reproductive organs of parents. This notion of pangenesis

[3]For instance, in the preface to the second edition of *The Descent of Man*, he emphasizes the relevance of "what I have called 'correlated' growth, meaning, thereby, that various parts of the organisation are in some unknown manner so connected, that when one part varies, so do others; and if variations in the one are accumulated by selection, other parts will be modified" (Darwin 1877: vi).

could explain the cross-generational transmission of environmentally induced characteristics which, for Darwin, was key in the appearance of novelties, as an evolutionary process different from adaptation through natural selection.[4] However, Darwin's pangenesis was not taken further, as explanations of the nineteenth and twentieth centuries progressively abandoned the theory of inheritance of acquired characteristics.

In contrast to this, the new rethinking of inheritance, which begins to emerge in the last third of the nineteenth century, considered that inheritance consists of the transmission of stable and fixed particulate elements, free from environmental influences. This idea, which precedes the rediscovery of Mendel's theory and begins to take form after the publication of the *Origin*, has been termed "hard inheritance" (Kampourakis 2017). It entailed a reconfiguration of the notion of inheritance which comprised Galton's statistical work and Weismann's proposal of the existence of a barrier between germ and somatic cells and excluded the inheritance of acquired characters from the Darwinian theory of evolution (Depew and Weber 2011; Noble 2021). The term "neo-Darwinism" was coined to designate the exclusion of inheritance of any kind of modification occurred during lifetime.

Challenges to gradualism gained a new dimension after the rediscovery of Mendelian laws of inheritance in 1900, as the gradualist character of Darwinian views seemed to conflict with the discrete character of Mendelian inheritance (Pigliucci 2009). Debates between Mendelians and the statistical approach to heredity of the Biometric School, which was based on Galton's work, lasted for some decades. However, recent studies of the early development of genetics suggest that the positions in the Mendelism-Biometry debate were not as clearly demarcated as it has often been stated, as the many intermediate positions among Biometricians were basically underestimated (Shan 2021). In addition, although most reconstructions of this episode in the history of genetics underlie the idea that Mendelism was superior to Biometry, this assumption has recently been called into question. For instance, Radick (2005) suggests that biometrician Weldon's views of the role of genes in heredity and evolution are closer to current views than previously noticed. Philosophically, the Mendelism-Biometry controversy concerns the nature of variation: whereas Mendelians like Bateson emphasized the significance of discontinuous variation, biometricians like Weldon insisted that relevant modifications are gradual. In any case, it is clear that Darwinian gradualism has not ceased to be an important topic of discussion on evolutionary ontology.

In sum, in addition to questioning gradualism, the new understanding of heredity led to a progressive shift away from the Darwinian conception of organisms as purposive entities organized in relationship to their environment, towards a view that emphasized the innate determination of organismic traits by genetic factors. This shift in the approach to inheritance had a major impact on evolutionary ontologies

[4] Mary Jane West-Eberhard has argued that the occurrence of large variants via environmental induction does not contradict Darwin's gradualism since they are the product of previous gradual variation (West-Eberhard 2008).

because it was now assumed that the foundations of the stability and organization of life are grounded in levels of organization below the organism.

3 From the Modern Synthesis to a Gene-Based View of Evolution

The development of population genetics advanced by figures such as Ronald Fisher, Sewall Wright, and J. B. S. Haldane during the 1920s and 1930s moved towards a statistical resolution of the Mendelism-Biometry dispute by formulating a series of mathematical models that would abstract continuous evolutionary change from the combination of small genetic elements. The success of this kind of theoretical project which imported methods based on statistical mechanics to biology was highly influential in the subsequent development of the Modern Synthesis, which would aim at unifying biology under evolutionary statistical terms, but it neglected the organismal level and diverted attention from wholes to particles, such as genes.

By contrast, already in the 1930s scholars such as Ludwig von Bertalanffy or Joseph Henry Woodger had called for the elaboration of theoretical foundations that were intended to unify biology as a science and orient empirical work around organizational tenets. They perceived the need for this theoretical orientation for a science that was developing on the basis of fragmented and confused ideas (see Etxeberria and Umerez 2006). For them, this work should be philosophical, on the one hand, and also oriented to developing a theory for biology inspired by the way theoretical physics is related to experimental physics. Betty Smocovitis (1996) contended that their demand for a "unifying principle" contributed to paving the way for the later unification via evolution carried out by the Modern Synthesis: "Evolution, purged of unacceptable metaphysical elements, would function as the phenomenon that could make biology an 'autonomous' science, at the same time that it served as the 'unifying principle' that Woodger and others had sought" (Smocovitis 1996: 114).

The explanatory framework of the Modern Synthesis was the result of a collective unifying project, pursued fundamentally during the 1930s and 1940s, with the aim of integrating the population genetic theory of evolution with naturalistic-systematic biology. New links were built between previously independent fields, and interactions between geneticists and taxonomists resulted in a series of publications that were pivotal for the development of the newly inaugurated research program. We may mention the works of Theodosius Dobzhansky, Julian Huxley, and Ernst Mayr, as the most influential contributions to the construction and early development of the synthetic theory of evolution.[5] Their main aim was to integrate the material basis of

[5] The major landmarks in the genesis of the synthetic theory of evolution are books by Dobzhansky (1937), Huxley (1942), and Mayr (1942). Dobzhansky's *Genetics and the Origin of Species* studied the genetics of natural populations focusing on geographic variations. He combined Wright's

evolution (i.e., the gene) and the mechanical cause of evolutionary change (i.e., natural selection), by building a mechanistic and materialistic science that would emancipate biology from physics and chemistry, and incorporate the peculiarities of evolving materials to explain organic phenomena (Smocovitis 1996). In the following decades, this unifying project would continue to develop, and other biological disciplines would also be reframed under this explanatory framework, such as paleontology, with the studies of George Gaylord Simpson, or botany, after the works of George Ledyard Stebbins (Bowler 2003; Smocovitis 2001).

However, despite the shared commitment to mechanize evolution, at least Huxley, Dobzhansky, and Mayr disagreed with an excessively reductionist approach. For instance, Dobzhansky refused to eliminate emergentism entirely, at least in the evolution of humans (Mayr 1991; Smocovitis 1996 and references therein), and Mayr recurrently expressed his disconformity with the relevance of population genetics for understanding evolution.[6] Whether they ultimately built a sufficiently comprehensive framework has been widely criticized by biologists and philosophers of biology (see Sect. 4). In contrast to these concerns, in the following decades a molecularized and gene-centered view of evolution derived from the premises of the Modern Synthesis and the research advances in different life science fields.

The molecularization of evolution resulted from the maturation of the population genetic theory of evolution and the development of molecular biology during the second half of the twentieth century. The discovery of the DNA molecule and the mechanisms of its replication and transmission grounded the hypothesized concept of the gene in a material structure, and the Central Dogma of Biology (Crick 1970) gave rise to an updated reformulation of the Weismann Barrier by restricting inheritance to transmission of DNA strains (Noble 2021), thus strengthening the reliance of the synthetic theory on hard inheritance. An adaptationist, gene-centric evolutionary biology was developed, departing from the development of the new science of molecular genetics and a rather restricted interpretation of Darwinism. Popular through Richard Dawkins' work on the selfish gene (Dawkins 1976, 1982), this theory grants agency to genes and subtracts it from organisms, which are

mathematical models with the populational approach of the Russian school to claim that morphological differences between populations have a genetic basis. Huxley not only coined the term "Modern Synthesis" in his book *Evolution: The Modern Synthesis* but also worked towards a synthesis of the evolutionary knowledge of the time, aiming at bringing together different biological disciplines from an integrative point of view. Mayr's *Systematics and the Origin of Species* explores the mechanisms of speciation and the effects of geographic variation and isolation. He also made major contributions to the history and philosophy of biology, especially through the concepts of the autonomy of biology.

[6]Mayr argued that the contribution of mathematical models to evolutionary biology overall was somewhat scarce, due to the "gross simplification" of the biological phenomena that these models required (Mayr 1959, in Provine 2004). Besides, Mayr criticized the role of population genetics within the theory of evolution on the basis of his (admittedly underinformed) idea that population geneticists disregarded phenomena of epistasis, i.e., the interaction between genes (see Rao and Nanjundiah 2011 for a historical review of the so-called Beanbag Dispute between Mayr and Haldane).

understood as "sets of relatively discrete adaptations" (Depew and Weber 2011: 94). The role of the organism is reduced to being merely a *vehicle* of genes, an intermediary between the genetic level, at which variation occurs at random, and the ecological level, in which natural selection acts. As an explanatory framework, it reduces evolution to allele modification via the action of natural selection, identifies genes with particular biological functions, and is applied to evolutionary phenomena at all levels: from the molecular to the social. As a consequence, it disregards phenomena occurring at the level of the organism, such as development (including constraints or biases to variation), organizational properties and dispositions, and environmental interactions (both with the inorganic milieu and with other organisms from the same or different species).

Hence, the current mainstream version of the theory of evolution is oriented towards a genetic explanation of variation, a project that has remained unchanged for the last seven or eight decades (Müller 2017). This form of explanation is believed to be *complete* in the sense that all evolution is assumed to be a direct consequence of changes in gene frequencies within the population at matter, while macroevolutionary patterns, speciation, variation of forms, etc. are seen as reducible to explanations in terms of relative allele frequencies.

Recently there have been new attempts to reexamine the history of the Synthesis, not only because some contents and fields were left out of the intended unification but also to review the epistemological and metaphysical commitments that first motivated it and aided to its development. In particular, although the synthesis was considerably successful at a sociological level, there were "internal cracks" at the conceptual synthesis (Delisle 2011; Gayon and Huneman 2019). The aim to view all living phenomena under the evolutionary prism importantly influenced the way in which biological ontologies were conceived, as the well-known motto "Nothing in Biology Makes Sense Except in the Light of Evolution" stated (Dobzhansky 1973). Admittedly, however, the organism was progressively displaced, and new kinds of *evolutionary individuals* started to conform to the ontology of biology.[7]

4 Two Kinds of Ontological Objections to the Modern Synthesis

This section summarizes two of the main ontological objections to the framework of the Modern Synthesis: one of them concerns structuralist objections to gradualism, while the other one has to do with biological ontologies at two different levels – the genetic and the organismic – and questions the nature, properties, and capabilities of different individuals and units of evolution.

[7]The task of reviewing biological individuality is beyond the scope of this work. Relevant references can be found in Godfrey-Smith (2013) and Pradeu (2016).

4.1 Gradualists and Structuralists (or Typologists)

On the first matter, philosopher Marjorie Grene criticized many of the philosophical positions developed by the Modern Synthesis. She defended nearly Aristotelian views against Mayr's (and other's) attacks to the "typological" thinking and views of species as individuals (Grene 1976, 1990; Honenberger 2015). Particularly, Grene (1959) examines the controversy between gradualist and discontinuist views when addressing the appearance of new forms: according to Simpson, evolution proceeds continuously, and new forms are the result of small cumulative adaptations; contrarily, Otto Schindewolf considers discontinuities in the fossil record to be fundamental and therefore focuses on the occurrence of new forms and types in evolution. Grene notes that the confrontation between the two scientists lies primarily in the fact that they focus on two different ontological properties of evolution: on the one hand, the continuity of life, which is addressed by Simpson's neo-Darwinian stance, and, on the other hand, the diversity of ordering principles that can be identified, which is the main target of Schindewolf's approach. Grene contends that those ordering principles cannot be fully reduced to the neo-Darwinian explanatory framework, since a mechanistic approach does not suffice for addressing the appearance of new characters. These evolutionary phenomena cannot be accounted for as being just the outcome of a reshuffling of the initial conditions but require a different logical stance. Hence, in order to fully address the nature of evolutionary change, both the logic of continuity and the logic of the emergence of discontinuities need to be considered.

A similar discomfort with the Synthesis was voiced by paleobiologists Niles Eldredge and Stephen J. Gould (1972) as well, who observed that new findings in the fossil series did not seem to match the predictions of the synthetic theory of evolution. They noted that the tempo of evolution at a paleontological scale does not occur continuously and gradually, as predicted by the synthetic models, but rather phenomena of stasis (stable maintenance of biological forms during important periods of time) alternate with moments of rapid diversification of forms. As the allopatric model of speciation – which relies on natural selection and migration (and often, also drift) as the only relevant factors in the evolution of species – cannot explain evolutionary stasis, these authors argue that other elements such as the self-regulation of development are required to understand evolution at geological scales. The theory of *punctuated equilibrium* tries to make sense of these phenomena.

A few years later, one of the most relevant pieces of criticism to the evolutionary view of the Synthesis was presented by Stephen J. Gould and Richard Lewontin in their renowned *Spandrels of San Marco and the Panglossian Paradigm* (Gould and Lewontin 1979). They confront one of the main pillars of the Modern Synthesis: its reliance on adaptation-based explanations of evolutionary events. They argue that the adaptationist program fails in distinguishing the present utility of a trait from the conditions or reasons of its evolutionary origination. According to them, organisms within the adaptationist framework are "atomized into 'traits' [that] are explained as structures optimally designed by natural selection for their functions" (Gould and

Lewontin 1979: 585). As opposed to this view, Gould and Lewontin argue that organisms are integrated entities whose *Baupläne* (or body plans) are "replete with constraints upon adaptation [. . .] that restrict possible paths and modes of change so strongly that the constraints themselves become much the most interesting aspect of evolution" (Gould and Lewontin 1979: 594). They propose an alternative way of understanding both organisms and their evolution in terms of structural integration and developmental constraints, yet still acknowledging the role of natural selection in the evolution of the living. Hence, the organization of living beings has to be understood as logically previous to the action of variation-introducing processing and natural selection itself, because the very organization of the organism shapes the framework of possibilities that are available for any organism.

4.2 Objections to the Genetic View of Evolution

The causal prioritization of the genetic level in biological explanations has long been criticized by developmental biologists. One of the first and more influential among them was Conrad Hal Waddington, who consistently challenged the gene-centered ontology of the Modern Synthesis on the basis of his empirical and theoretical investigations on organismal evolution. In particular, he worked on the developmental processes that shape phenotypes and highlighted the neglect, within the synthetic theory, of the complex processes linking the genetic and phenotypic levels, which he called the *epigenetics* of the organism (Waddington 1957: 13). According to his proposal, phenotypes resulting from mutations occurring at the genetic level are not random, as postulated by the proponents of the genetic theory of evolution, but they result from the interaction of genetic and nongenetic factors which are *canalized* by the intrinsic properties of the developmental processes (Waddington 1953, 1959, 1968). This preoccupation with the constructive role of development reflects Waddington's conviction that the processes and phenomena that take place at the level of the organism are fundamental for biological ontology.

Furthermore, critics of the Synthesis have claimed that a process of *hardening* took place after the formulation of the Modern Synthesis and during its maturation throughout the 1940s and 1950s. According to Gould, over the decades following the publication of the inaugural texts of the Synthesis, a more adaptationist version of it was gradually developed (Gould 1983). The first version of the synthesis, he argued, was driven primarily by the methodological concern that there is no independent genetic mechanism to explain large-scale evolution, whatever the cause of change is (natural selection, drift, mutations). Therefore, any theory of evolutionary change could be accepted as long as it was based on known Mendelian genetics. This is the reason why one of the main challenges for the builders of the Synthesis were the processes of speciation, which ought to be explained by the very same mechanisms that were at the time described in laboratory and local populations. In contrast, during the hardening phase, the Modern Synthesis became increasingly restrictive, eventually considering that gradual, cumulative natural selection leading to

adaptation was the only mechanism by which organisms could evolve.[8] For its part, the selfish gene theory introduces an innovative consequence: organisms are no longer the units of evolution as the builders of the Modern Synthesis had considered (Svensson 2021). The understanding of genes as entities that transmit evolutionary functions radically alters the ontology of evolution. The group selection debates of George C. Williams (1966) and Richard Dawkins (1976) played an important role in transforming evolutionary ontology by shifting the focus from organisms to genes in a reductionist and deterministic move that was influenced by sociobiology and behavioral ecology and had enormous implications for both the scientific and the public understanding of the nature of life (Svensson 2021).

5 The (Re)Introduction of Developmental and Ecological Aspects into Evolutionary Biology

The developmental and ecological dimensions of organisms, neglected by the Modern Synthesis, are reintroduced by current evolutionary studies. The Extended Synthesis[9] adopts an organismic perspective in its ontology in those two basic dimensions. Firstly, it assumes that the morphologies of organisms – their form and structure – are constructed by inherent, material, and self-organizing processes, in dialogue with the environment, not "designed" solely by adaptive demands. Thus, as regulation and feedback dynamics characterizes developmental systems, *internalism* is sometimes underlined in contrast to the *externalism* characteristic of the Modern Synthesis (Alberch 1989; Nuño de la Rosa and Etxeberria 2009). Secondly, the Extended Synthesis recognizes that environmental conditions themselves are largely structured by organisms, whose causal constructive role is agential as it follows purposeful behavior. The relevance of environmental factors is emphasized here as opposed to an excessive focus on the internal genetic perspective. In this ontology, the organism is the consequence of a historical process that endures from conception to death and in which the gene, the environment, chance, and the organism as a whole are continuously involved (Lewontin 1983).

This change of perspective affects also a very influential principle within the Modern Synthesis concerning the difference between proximate and ultimate causes in biology: according to this distinction, proximate causes answer *how* questions characteristic of functional biology, whereas ultimate causes are evolutionary and concern *why* questions (Mayr 1961). For some developmentalists, it is not enough to

[8] Gould portrays this tendency towards a stricter adaptationism in the successive changes undergone by the different editions of the main books of the Modern Synthesis (Gould 1983).

[9] The term "Extended Evolutionary Synthesis" was coined by Massimo Pigliucci (2007) to refer to the effort of unifying the theoretical framework of the Modern Synthesis with a theory of forms that would include concepts such as "evolvability, phenotypic plasticity, epigenetic inheritance, complexity theory, and the theory of evolution in highly dimensional adaptive landscapes" (Pigliucci 2007: 2743).

consider gene mutation and recombination to study variation: proximate causes, which are active in phenotype building, are also required. Hence, important drivers of phenotypic diversity, such as developmental bias or phenotypic plasticity, need to be integrated into the existing theory of evolution (Brigandt 2015, 2020; Brown 2022). For other scholars, biology should do away with the proximate-ultimate distinction and replace it with a notion of reciprocal causation in which "causation is perceived to cycle through biological systems recursively" (Laland et al. 2013: 719).

5.1 The Generative Role of Development

The "developmental perspective" emphasizes the role of development for evolution, since the appearance and alteration of structures, parts, features, etc. depend on the possibilities, constraints, and burdens of organized systems throughout their development. It is for this reason that evolution cannot be understood unless the development, organization, and structure of organisms are taken as fundamental pillars for evolutionary ontology. The study of these processes involves morphological and physiological research from a systemic and non-reductionist approach: Evolutionary Developmental Biology – or Evo-Devo – brings development and evolution together in this task.

Evo-Devo's observations around the fact that developmental processes can constrain phenotypic variation conflict with the standard ontology of the Modern Synthesis in that some forms are more likely than others in evolution. In fact, studies on developmental constraints suggest that phenotypic variation may be canalized and directed toward functional types by developmental processes (Alberch 1982; Brigandt 2015, 2020; Brown 2022). A causal plurality is therefore assumed in the determination of living systems, emphasizing the importance of developmental processes and the primacy of structure.

Pere Alberch, one of the precursors of Evo-Devo, emphasized the interactive nature of developmental processes, so that the complex interactions established between genes and other factors, such as cellular properties or tissue geometries, give rise to the processes of morphogenesis (Alberch 1991; Etxeberria and Nuño de la Rosa 2021). The study of how developmental processes evolve is based on the premise that the relationship between phenotype and genotype is not merely a statistical correlation, but that developmental rules govern the generation of phenotypes: they are a determining factor in evolutionary processes (Alberch 1982). The mapping of the complex processes at work in the interplay between genes and phenotypes has become a significant component of the current Evo-Devo research agenda. These studies on genotype-phenotype maps have their roots in twentieth-century developmental biology: to convey the complexity and relational character of these processes, Waddington (1957) introduced the *epigenetic landscape*, consisting in a multidimensional scheme with valleys and peaks, which models the complex interaction between genetic and environmental factors during development. The fact

that the relationship between genotype and phenotype is nonlinear both at the lower organizational level (protein transcription) and at the higher levels (developmental or morphological) suggests that development is an element with relevant evolutionary consequences in the appearance of new traits. Hence, Evo-Devo succeeds in acknowledging both internalist and externalist ontological positions with respect to the construction and evolution of organisms while bringing proximal and ultimate causes closer together.

5.2 Developmental Plasticity

One of the most widely studied phenomena in Evo-Devo is the influence of developmental plasticity in evolution. Thanks to this property, which allows an organism to alter its biochemistry, morphology, physiology, and/or behavior in response to its surrounding environmental conditions and is purportedly pervasive in highly evolved species, individual organisms can develop differentially in the face of external influences (West-Eberhard 2003). These environmentally induced variants involve a reversal in the causal order postulated by the Modern Synthesis, for which the genotype contains all the information necessary to generate a specific phenotype, and thus, phenotypic changes can only occur as a consequence of alterations in the genetic material. An alternative form of evolution is postulated in which phenotypic changes precede changes in the genotype (Levis et al. 2018; Levis and Pfennig 2019; West-Eberhard 2003). Plasticity-driven evolution is associated with the phenomena of phenotypic accommodation and genetic assimilation, which result from processes for the fixation of variants introduced by the environment, and rely on the systemic and organizational features of the organism (Pigliucci et al. 2006; Waddington 1953, 1959). Furthermore, the role of developmental plasticity in macroevolutionary changes is highlighted as being associated with the phenomenon of homoplasy and with the rapid divergence of phenotypic lineages (West-Eberhard 2003).

Recent plasticity-first accounts of trait evolution find that "a plastic response to environmental change precedes and enables adaptive change, compared to a genes-first mode, in which selection acts on new genetic variants first and changes in plasticity are secondary" (Paaby and Testa 2021: 1078). There have been different ways of incorporating plasticity into evolutionary studies; recent work has pointed out that idealizations of the relationships among the three components of evolutionary explanations – namely, variation, heredity, and selection – greatly hinder the understanding of the significance of plasticity (Uller et al. 2020). In fact, a proper account of plasticity should not consider the three of them as fully separable due to the influence of developmental causes. Consequently, advances in the systemic, organizational, and multi-level perspectives of organisms entail a reconsideration of inheritance, so that factors other than genetic are included as elements that are transmitted between generations. The notion of Extended Inheritance includes epigenetic, behavioral, and symbolic systems of inheritance, so that environmentally

induced factors are often actively conveyed in both vertical and horizontal dimensions (Jablonka and Lamb 2005; see also Bonduriansky and Day 2020).

5.3 Organisms as Ecological Agents

The ecological agential perspective of organisms involves considering how their activities shape the world in which they live and determine the selective pressures to which they (and their offspring) are exposed to. The agential perspective entails that rather than passive units molded by natural selection, organisms construct their environments, which affects the course of their own evolution. In this fashion, Niche Construction Theory elaborates the relationship between the organism and its environment as an active role. This theory was developed from the late 1980s onwards mainly by John Odling-Smee, Kevin Laland, and Marcus Feldman; according to their proposal, organisms modify their environment actively and non-randomly. In this way, it is the organisms of the population in question themselves who alter their niche and construct the environmental conditions to which subsequent individuals will be exposed, thus co-directing their own evolution and influencing the evolution of other species (Laland et al. 2016). As a consequence, niche construction processes can lead to the fixation of alleles that would otherwise prove deleterious and can facilitate the maintenance and survival of organisms in environments with adverse conditions (Laland et al. 1999). In this way, the environment itself has been sometimes regarded as being itself inheritable.

6 The Role of the Organism: Organicism, Evolution, and the Extended Synthesis

These developments lead to a re-signification of the theoretical role of the organism in both biology in general and evolutionary biology in particular. New insights on the developmental, systemic, and relational aspects of organisms are taken into consideration, as a consequence of which the form, structure, organization, and dynamics of the organism become a central element in characterizing organic matter. This tendency gave rise to a new "organism-centered biology" (Baedke 2019; Etxeberria and Umerez 2006).

As mentioned in Sect. 3 above, in the early twentieth century, there were attempts to unify biology according to organizational principles based on systemic views of the organism. The notion of organism was crucial to these projects in theoretical biology in the 1920s and 1930s which conceived of it as the fundamental object of study in the scientific research on life. It was carried out by drivers such as Conrad Hal Waddington, Joseph Henry Woodger, or Joseph Needham in Cambridge and Ludwig von Bertalanffy and Paul Weiss in Vienna. The organism was considered a

central concept in biology, since it was understood that its properties did not depend on the peculiarities of its components, since they were systemic (Etxeberria and Umerez 2006). These advances in biology were overshadowed by the rise of molecular biology and the genetic "hardening" of the Modern Synthesis (see Sect. 4).

Today, the growing emphasis on the systemic and organizational properties of the organism and the attention being paid to the generative properties of development support a return to an organism-based approach to evolution. Daniel Nicholson has suggested that the revival in the interest in organisms occurring during the last two decades is due to the consideration that the Modern Synthesis is not able to offer a satisfactory understanding of the phenomena of biological evolution, since the systemic properties of the organisms are neglected. This situation thus motivates the search for alternatives that take into account the complexity of organisms in order to overcome the reductionist perspective introduced by molecular biology. Finally, the renewed interest in considering the question about the nature of life brings the concept of the organism into question (Nicholson 2014).

Our conceptual review of the ontologies covered by the two syntheses supports the thesis that the organism is particularly relevant for understanding the evolution of forms and functions, at least in the context of Extended Synthesis; this point has also been suggested by Fábregas-Tejeda and Vergara-Silva (2018b). In a further note, Niles Eldredge singles out the organismic level as a special one in his Hierarchical Theory of Evolution (Eldredge 1985). He postulates two different hierarchical scales in the ontology of the living: one related to the development, retention, and modification of the information contained in the genome, consisting of codons, genes, organisms, demes, species, and monophyletic taxa (replicators), and another one constituted by ecological individuals, which include proteins, cells, organisms, populations, communities, and regional biotic systems (interactors). The organism level occurs in the two hierarchies, something that Eldredge interprets as a hint that the organism level constitutes some sort of connection between the two hierarchies, thus relating the informational dimension to the ecological one.

Furthermore, other authors stress that this linking of the two hierarchies may have even a greater depth than what Eldredge had assumed (Ruiz-Mirazo et al. 2000). They highlight that "if living beings such as bacteria and vertebrates, which are functionally and organizationally so different, are included under the label 'organism' it is because they share certain features that define them as units in both hierarchies" (Ruiz-Mirazo et al. 2000: 229–230). Accordingly, they regard the notion of organism as *primitive* concept in biology, irreducible to its components, processes, or mechanisms and propose an account of the organism based on four interrelated elements: (i) *individuality*, considered as the characteristic of organisms (and not of other categories) to be individual in the sense of being unique both due of the genetic material they possess and because they are the result of their own particular history, development, and interactions; (ii) the *organization* of the system in such a way that the organism is conceived as an entity whose parts contribute to the constitution of the system. It is considered that parts and wholes require different explanations, there being an intricate set of dynamic relationships between different

levels of organization; (iii) *autonomy*, which refers to the relative independence of the system with respect to its environment. This form of autonomy is a particular kind of self-organization, in which the components that constrain and define the organization of the system are produced by the system itself, this being the property that differentiates the organization of living beings from the organization of machines[10]; and (iv) *reproduction*, which provides a criterion for differentiating individual organisms from individual colonies or societies. Reproduction is considered as the process by which an organized system is capable of generating a similar self-organized autonomous system. This four-dimensional notion of the organism offers a general view of the particularities of organisms, based on fundamental aspects of life that are exclusive of them and emphasizes their ontological and epistemic saliency for biology and evolutionary biology in particular.

However, current organismic biology faces a series of difficulties which, according to Jan Baedke's historical examination, are common to its early-twentieth-century antecedents in organismic biology (Baedke 2019). Both approaches share motivations and theoretical components, including the focus on responsiveness to the environment, plasticity, the processes underpinning organizational robustness, and organism-environment interactions. Some of the shared challenges relate to the difficulties raised by emphasizing the causal relevance of the environment in the construction and evolution of organisms, as well as their capacity to modify environmental characteristics – that is, the reciprocity between both elements – so that the border between organism and environment is blurred (Baedke et al. 2021). Moreover, boundaries among individual organisms appear to be disintegrating, in cases of symbiosis constituting holobionts (Suárez 2018). These considerations alert on the requirement of further conceptual developments to clarify unsettled issues about borderline cases that challenge the nature and properties of the organism as well as its relationships with its environment and with other organisms. The philosophy of the organism is thus a rich and promising enterprise in which many questions, such as causal reciprocity, emergentism, body-mind questions, reproductive relationships, etc., remain open.

The development of the research agendas necessary to carry out this work (including Evo-Devo, Niche Construction Theory, and organismic biology) is envisaged as a task for the so-called Extended Synthesis, which has often been interpreted as successive expansions of the synthetic theory of evolution both in Kuhnian (Pigliucci 2009) and Lakatosian terms (Pievani 2012), seeking for a unification of scientific concepts and disciplines. This enterprise has recently been criticized by Fábregas-Tejeda and Vergara-Silva, for whom the unifying ideal of evolutionary biology can be traced back to logical empiricism (Fábregas-Tejeda and

[10]Not just any form of organization accounts for the ontology of autonomous organisms. It is required that the components of the system have an active role in the maintenance and functioning of the system, so that both the constituents and the interactions among these components need to be considered from a material and historical stand. Hence, the organism cannot be understood except in relation to the processes and constituents that make it up (Moreno et al. 2008; Ruiz-Mirazo et al. 2000).

Vergara-Silva 2018a). In contrast to this, they propose to conceive the new research framework under the metaphor of a dynamic network of interrelated models and representations, instantiated in a diversity of practices, whose nodes may change their relative position, centrality, and connections in time. The components of this network are articulated around *epistemic goals* (Brigandt 2010) or *problem agendas* (Love 2010), which motivate the scientists' endeavors. This model better captures the complexity and dynamic nature of scientific practices; instead of being constructed as unitary bodies of explanation, all particular elements are coordinated and articulated within a dynamic and complex network of interrelations. Then, the two syntheses would not be considered as separate and individualized bodies of knowledge or as successive strata in the development of evolutionary biology.

We have argued throughout this article that the ontologies compromised by the characteristic projects of each of the two syntheses constitute a fundamental discordance between their different approaches to biological evolution. Our analysis shows that the two syntheses rely on different ontological commitments: although the Extended Synthesis constitutes a more pluralistic framework, in which works examining phenomena at different ontological levels may coexist, the Modern Synthesis had become increasingly reductionist and ontologically monist. Therefore, the project of the Extended Synthesis needs to be considered as a non-cohesive cluster in a dynamical network-like array of models, in which the trading and exchange of models and concepts between disciplines are the rule instead of being the exception. Accordingly, this array of models may have a better flexibility to be inclusive and to potentially shift among different ontological levels without necessarily aspiring to integrate them all in a single unified framework. We found that the examination of the ontologies adopted by the different projects has helped to enrich and complete previous epistemological analyses of the commitments of the two Syntheses.[11]

7 Conclusions

This chapter is articulated as an ontological examination of the role of organisms in evolutionary theory. Our overview reviews different conceptions of the organism throughout the history of evolutionary biology; it highlights that the organism was first assumed without much questioning in Darwin's work, then fell into crisis after some late developments of the Modern Synthesis adopted a gene-centric view of life, and, more recently, it is once again considered as a major theoretical concept in biology within the Extended Synthesis.

During its maturation, the Modern Synthesis progressively excluded the organism as a theoretical concept of biology as it was involved in a defense of evolution-

[11] We thank our anonymous reviewer for helping us to formulate this suggestion.

ary principles which were committed to gradualism with respect to inheritance and population thinking in the understanding of variation. Those features shaped some of the gene-centered, adaptationist ontological commitments of the Modern Synthesis. We contend that the most noticeable discordances between the Modern Synthesis of evolution and proposals for an Extended Synthesis concern their ontological commitments with respect to the role of organizational and developmental mechanisms in evolution and the processes of life. Some of the research problems being explored by the Extended Synthesis can be conceived as a "return of the organism," conceptualized in organizational and developmental terms. On the more recent recovery of the organism, we have pointed out that two important dimensions have played a central role. One of them concerns the growing attention paid to the development of organisms, which implies that developmental processes from a systemic perspective need to be considered to study variation. The second dimension has to do with the consideration of the ecological relationships of organisms in which they appear as subjects of evolution and as active agents in their environments.

A strong, organizational, relational, and agential notion of the organism becomes inevitable for understanding many phenomena, without which evolutionary biology is incomplete. Moreover, organismic assumptions – such as addressing multiple layers of reciprocal causality, the active role of the organism as an agent, or the striking relevance of developmental processes in constraining the course of evolution – are incompatible with the neo-Darwinian enterprise. In relation to the historical relations within evolutionary biology, we argue that rather than the successive unifications or expansions of the theoretical framework that are usually presumed, it is the scientific activity of the actors involved in evolutionary explanations that displays a plural set of research questions and gives rise to a network-like array of models and practices that constitute evolutionary biology, whose epistemological aspects are importantly influenced by the ontologies that different theories may compromise.

In sum, we have examined discordances between the two syntheses from an ontological standpoint focused on organisms. We have argued that whereas the framework of the Modern Synthesis became increasingly reductionist and monist, the one of the Extended Synthesis is constituted by a pluralist array of models able to accommodate different ontological levels, although the organism certainly constitutes a crucial element at the crossroads of many other significant ontological aspects because of its flexibility and potential inclusiveness.

Acknowledgements DCG and AEA take part in Funding for Research Groups of the Basque Government [IT1668-22] and in the *Metaphysics of Biology* MICINN Research Project [PID2021-127184NB-I00]. AEA is also part of the *Outonomy* MICINN Research Project [Ref PID2019-104576GB-I00]. DCG has pre-doctoral contract from the UPV/EHU (PIF-2020). We thank our editors for their kindness and our anonymous referees for the very generous comments and advise they did to the previous versions of this chapter.

References

Alberch P (1982) Developmental constraints in evolutionary processes. In: Bonner JT (ed) Evolution and development. Springer, Berlin Heidelberg, Berlin, pp 313–332
Alberch P (1989) The logic of monsters: evidence for internal constraint in development and evolution. Geobios 22:21–57
Alberch P (1991) From genes to phenotype: dynamical systems and evolvability. Genetica 84(1): 5–11
Baedke J (2019) O organism, where art thou? Old and new challenges for organism-Centered biology. J Hist Biol 52(2):293–324
Baedke J, Fábregas-Tejeda A, Prieto GI (2021) Unknotting reciprocal causation between organism and environment. Biol Philos 36(5):48
Bonduriansky R, Day T (2020) Extended heredity: a new understanding of inheritance and evolution. Princeton University Press, Princeton
Bowler PJ (2003) Evolution: the history of an idea. University of California Press, Berkeley and Los Angeles
Brigandt I (2010) Beyond reduction and pluralism: toward an epistemology of explanatory integration in biology. Erkenntnis 73(3):295–311
Brigandt I (2015) From developmental constraint to evolvability: how concepts figure in explanation and disciplinary identity. In: Love AC (ed) Conceptual change in biology: scientific and philosophical perspectives on evolution and development. Springer, Netherlands, pp 305–325
Brigandt I (2020) Historical and philosophical perspectives on the study of developmental bias. Evol Dev 22(1–2):7–19
Brooks DR (2000) The nature of the organism: life has a life of its own. Ann N Y Acad Sci 901(1): 257–265
Brown RL (2022) Structuralism and adaptationism: friends? Or foes? Semin Cell Dev Biol. https://doi.org/10.1016/j.semcdb.2022.02.022
Cheung T (2010) What is an "organism"? On the occurrence of a new term and its conceptual transformations 1680-1850. Hist Philos Life Sci 32(2–3):155–194
Crick F (1970) Central dogma of molecular biology. Nature 227(5258):561–563
Darwin C (1877) The descent of man, and selection in relation to sex. John Murray, United Kingdom
Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
Dawkins R (1982) The extended phenotype. Oxford University Press, Oxford
Delisle RG (2011) What was really synthesized during the evolutionary synthesis? A historiographic proposal. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci 42(1):50–59
Depew DJ, Weber BH (2011) The fate of Darwinism: evolution after the modern synthesis. Biol Theory 6(1):89–102
Dobzhansky T (1937) Genetics and the origin of species. Columbia University Press, Columbia
Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teach 35(3):125–129
Eldredge N (1985) Unfinished synthesis: biological hierarchies and modern evolutionary thought. Oxford University Press, Oxford
Eldredge N, Gould SJ (1972) Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf TJ (ed) Models in paleobiology. Freeman Cooper & Company, San Francisco, pp 82–115
Etxeberria A, Nuño de la Rosa L (2021) Pere Alberch (1954-1998). In: Nuño de la Rosa L, Müller G (eds) Evolutionary developmental biology—a reference guide. Springer
Etxeberria A, Umerez J (2006) Organismo y Organización En la Biología Teórica ¿Vuelta Al Organicismo? Ludus Vitalis 14(26):3–38
Fábregas-Tejeda A, Vergara-Silva F (2018a) The emerging structure of the extended evolutionary synthesis: where does evo-devo fit in? Theory Biosci 137(2):169–184

Fábregas-Tejeda A, Vergara-Silva F (2018b) Hierarchy theory of evolution and the extended evolutionary synthesis: some epistemic bridges, some conceptual rifts. Evol Biol 45(2):127–139
Gayon J, Huneman P (2019) The modern synthesis: theoretical or institutional event? J Hist Biol 52(4):519–535
Godfrey-Smith P (2013) Darwinian individuals. In: Bouchard F, Huneman P (eds) From groups to individuals: evolutionand emerging individuality. The MIT Press, Cambridge, MA, pp 17–36
Gould SJ (1982) Darwinism and the expansion of evolutionary theory. Science 216(4544):380–387
Gould SJ (1983) The hardening of the modern synthesis. In: Grene M (ed) Dimensions of Darwinism. Cambridge University Press, Cambridge, pp 71–93
Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London. Series B. Biol Sci 205(1161):581–598
Grene M (1959) Two evolutionary theories. Br J Philos Sci 9(35):110–127. 185–193
Grene M (1976) Aristotle and modern biology. In: Grene M, Mendelsohn E (eds) Topics in the philosophy of biology. Springer, Dordrecht, pp 3–36
Grene M (1990) Evolution, "typology" and "population thinking". Am Philos Q 27(3):237–244
Honenberger P (2015) Grene and Hull on types and typological thinking in biology. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci 50:13–25
Huxley JS (1942) Evolution: the modern synthesis. George Allen & Unwin, London
Jablonka E, Lamb MJ (2005) Evolution in four dimensions: genetic, epigenetic, Behavioral, and symbolic variation in the history of life. MIT Press, Massachusetts
Kampourakis K (2017) Making sense of genes. Cambridge University Press, Cambridge
Laland KN, Matthews B, Feldman MW (2016) An introduction to niche construction theory. Evol Ecol 30:191–202
Laland KN, Odling-Smee FJ, Feldman MW (1999) Evolutionary consequences of niche construction and their implications for ecology. Proc Natl Acad Sci 96(18):10242–10247
Laland KN, Odling-Smee J, Hoppitt W, Uller T (2013) More on how and why: cause and effect in biology revisited. Biol Philos 28(5):719–745
Levis NA, Isdaner AJ, Pfennig DW (2018) Morphological novelty emerges from pre-existing phenotypic plasticity. Nat Ecol Evol 2(8):1289–1297
Levis NA, Pfennig DW (2019) Phenotypic plasticity, canalization, and the origins of novelty: evidence and mechanisms from amphibians. Semin Cell Dev Biol 88:80–90
Lewontin RC (1983) The organism as the subject and object of evolution. Scientia 77(18)
Love AC (2010) Rethinking the structure of evolutionary theory for an extended synthesis. In: Pigliucci M, Müller G (eds) Evolution—the extended synthesis. MIT Press, Massachusetts, pp 403–441
Mayr E (1942) Systematics and the origin of species. Columbia University Press, Columbia
Mayr E (1959) Where are we? Cold Spring Harb Symp Quant Biol 24:1–14
Mayr E (1961) Cause and effect in biology: kinds of causes, predictability, and teleology are viewed by a practicing biologist. Science 134(3489):1501–1506
Mayr E (1991) One long argument: Charles Darwin and the genesis of modern evolutionary thought. Harvard University Press, Harvard
Meloni M (2016) Political biology: science and social values in human heredity from eugenics to epigenetics. Palgrave Macmillan UK, London
Moreno A, Etxeberria A, Umerez J (2008) The autonomy of biological individuals and artificial models. Biosystems 91(2):309–319
Müller GB (2017) Why an extended evolutionary synthesis is necessary. Interface Focus 7: 20170015
Nicholson DJ (2014) The return of the organism as a fundamental explanatory concept in biology. Philos Compass 9(5):347–359
Noble D (2021) The illusions of the modern synthesis. Biosemiotics 14(1):5–24

Nuño de la Rosa L, Etxeberria A (2009) Partes y funciones en el desarrollo y la evolución. Hacia un darwinismo sistémico. In: Dopazo H, Navarro A (eds) Evolución y Adaptación: 150 años después del Origen de las Especies. Obrapropia, Valencia, pp 465–474

Paaby AB, Testa ND (2021) Developmental plasticity and evolution. In: Nuño de la Rosa L, Müller GB (eds) Evolutionary developmental biology: a reference guide. Springer International Publishing, pp 1073–1086

Pievani T (2012) An evolving research programme: the structure of evolutionary theory from a Lakatosian perspective. In: Fasolo A (ed) The theory of evolution and its impact. Springer, Milan, pp 211–228

Pigliucci M (2007) Do we need an extended evolutionary synthesis? Evolution 61(12):2743–2749

Pigliucci M (2009) An extended synthesis for evolutionary biology. Ann N Y Acad Sci 1168:218–228

Pigliucci M, Murren CJ, Schlichting CD (2006) Phenotypic plasticity and evolution by genetic assimilation. J Exp Biol 209(12):2362–2367

Pradeu T (2016) Organisms or biological individuals? Combining physiological and evolutionary individuality. Biol Philos 31(6):797–817

Provine WB (2004) Ernst Mayr: genetics and speciation. Genetics 167(3):1041–1046

Radick G (2005) Other histories, other Biologies. Royal Institute of Philosophy Supplements 56: 21–47

Rao V, Nanjundiah V (2011) J. B. S. Haldane, Ernst Mayr and the beanbag genetics dispute. J Hist Biol 44(2):233–281

Ruiz-Mirazo K, Etxeberria A, Moreno A, Ibáñez J (2000) Organisms and their place in biology. Theory Biosci 119(3–4):209–233

Shan Y (2021) Beyond Mendelism and biometry. Stud Hist Philos Sci Part A 89:155–163

Smocovitis VB (1996) Unifying biology: the evolutionary synthesis and evolutionary biology. Princeton University Press, Princeton

Smocovitis VB (2001) G. Ledyard Stebbins and the evolutionary synthesis. Annu Rev Genet 35(1): 803–814

Sober E, Orzack SH (2003) Common ancestry and natural selection. Br J Philos Sci 54(3):423–437

Suárez J (2018) The importance of symbiosis in philosophy of biology: an analysis of the current debate on biological individuality and its historical roots. Symbiosis 76(2):77–96

Svensson E (2021) The structure of evolutionary theory: beyond neo-Darwinism, neo-Lamarckism and biased historical narratives about the modern synthesis. In: Dickins TE, Dickins JA (eds) Evolutionary biology: contemporary and historical reflections upon core theory. Springer International Publishing, Cham

Uller T, Feiner N, Radersma R, Jackson ISC, Rago A (2020) Developmental plasticity and evolutionary explanations. Evol Dev 22(1–2):47–55

Waddington CH (1953) Genetic assimilation of an acquired character. Evolution 7(2):118–126

Waddington CH (1957) The strategy of the genes: a discussion of some aspects of theoretical biology. Allen & Unwin, London

Waddington CH (1959) Canalization of development and genetic assimilation of acquired characters. Nature 183(4676):1654–1655

Waddington CH (1968) Towards a theoretical biology. Nature 218(5141):525–527

Walsh DM (2015) Organisms, agency, and evolution. Cambridge University Press, Cambridge

West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, Oxford

West-Eberhard MJ (2008) Toward a modern revival of Darwin's theory of evolutionary novelty. Philos Sci 75(5):899–908

Williams GC (1966) Adaptation and natural selection: a critique of some current evolutionary thought. Princeton University Press, Princeton

Tree Thinking and the Naturalisation of Language

Antonio Danese

Abstract This paper reconstructs Darwin's reflections on the evolution of language to underline how his conclusions are fundamental to legitimating an epistemological extension of naturalistic explanations of the origin of language. This extension allows us to overcome difficulties generated by the explanatory approach of evolutionary psychology (Pinker S, Bloom P, Behav Brain Sci 13:707–727, 1990).

Following the pluralistic approach of Darwinian tree thinking, I will try to show that if we want to give contemporary philosophers an epistemological justification for naturalising language, we have to conceive it as a mosaic of interactive components (descended larynx, vocal imitation, protolanguage made up of expressions and gestures etc.) with their own ancient or more recent evolutionary histories. Each component is necessary, but none are central or the most important. Each requires the integration of multiple explanatory processes (natural selection, exaptation, group selection, cultural evolution etc.), but no evolutionary mechanism alone is sufficient. Furthermore, each has to be studied phylogenetically.

Keywords Evolutionary psychology · Language · Evolution · Darwin · Exaptation

1 Introduction

This contribution aims to examine Charles Darwin's thought about the origin of speech and historical languages in relation to the previous and contemporary historical-philosophical and scientific reflections, in order to point out the fundamental methodological outlines for the naturalisation of language.

With the publication of the writings of Charles Darwin, philosophers were forced to reformulate their methods of rational reflection regarding the origins of language. According to Darwin the chief distinction between humanity and the lower animals was not the comprehension of articulate sounds, the articulation or the ability to

A. Danese (✉)
University of Padua, Padua, Italy
e-mail: antonio.danese@studenti.unipd.it

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_10

connect sounds with ideas: despite the fact that articulate language is peculiar to man, the ability of expressing meanings through the use of inarticulate cries, gestures, and the movements of muscles of the face is a trait in common with the lower animals. What makes human unique is the 'almost infinitely larger power of associating together the most diversified sounds and ideas' (see Darwin 1874: 85).

In order to explain the nature of language, the author proposed a methodological pluralism for a coherent understanding of the biological and cultural processes at the basis of the emergence of this faculty through a naturalistic explanation even for the most complex phenomena from the point of view of their symbolic and cultural expression.

Specifically, the need to introduce a naturalistic perspective has led philosophy to build a dialogue with linguistics, evolutionary biology, ethology, population genetics, palaeoanthropology, neurosciences and evolutionary psychology. The aim of this naturalistic perspective is to reconstruct the history of this faculty and the impact of biological constraints and sociocultural conditioning to establish thresholds, limits, points of contact and continuities with animal intelligence and vocality. This would allow us to understand how much language and mental structures have been shaped reciprocally.

Furthermore, the historian of science, whose point of view permeates this paper, has been tasked with reordering the interpretations and models of the past, finding the current developments and discovering if there are epistemological points of contact that allow them to be placed in a relationship of continuity or if there have been gaps that have led to total reconfigurations.[1]

The first part of this paper describes some of the main features of evolutionary psychology. After this preliminary section, the paper is divided into three parts.

The first concerns how Charles Darwin tried to explain the evolution of language through the same concepts he used for the evolution of species. Then I intend to show the limits of this approach to the study of language, and at the end of this part, I discuss two main philosophical consequences: Darwinian perspective allows us to overcome essentialism and recognise that on the one hand there are not languages more or less perfect and on the other hand the differences in languages can't be a stepping-stone for racist views.

The second portion of the paper concerns Darwinian definition of language, which involves the hybrid nature of this faculty: I explain why the concept of continuity is useful in constructing evolutionary explanation of language; subsequently, the paper explores the concept of imitation, and then it centres on the issue of coevolution. Focusing on these concepts is the occasion for examining the role of the comparative study of human and animal faculties, the coadaptation of language and brain and why Darwinian hypotheses are still highly topical.

[1] This methodological need for linking different epistemological fields to dissect the problem of language has encountered various obstacles over the years. The first and most influential occurred in 1886, when in Article 2 of the statute of the Société de Linguistique of Paris, it was established: 'La Société n'admet aucune communication concernant, soit l'origine du langage, soit la création d'une langue universelle'. A ban was then also reaffirmed by the Philological Society of London in 1872.

Then the paper turns to another debate that currently occupies the discussion about the origin of language: is Darwinian approach to the evolution of language able to cross the threshold of the origin of syntax? At the end of this second part, I discuss subjects which are worthy of philosophical attention: the differences between human language and animal communication systems, the idea that language is an innate capacity and the empiricist assumption according to which language is for expressing thoughts.

The third part describes the tree thinking and its role in the naturalistic approach to language. Since adaptationists accept the guiding idea that natural selection is overwhelmingly the most important cause of evolution of language, I move away from this conception, and I discuss how the system of ancestor/descendant relationship exhibited by the tree of life can permeate nonadaptive explanation for the evolution of language. Tree thinking and Fitch's consilience approach will allow us to solve the difficulties generated by the selective gradualism of the evolutionary psychology approach to language.

In the conclusion, I will sum up the meaning of the naturalisation of language meant as multimodal naturalistic approach based on the recourse to Darwinian evolution.

2 Evolutionary Psychology

In recent years, a naturalistic approach to language that indiscriminately resorts to the concept of adaptation has emerged. According to evolutionary psychology (Pinker and Bloom 1990), natural selection has designed language through gradual adaptive change in an environment of evolutionary adaptedness.

Steven Pinker and Paul Bloom describe language development in adaptationist terms, giving the selection an absolute power capable of leading to the origin and gradual implementation of language, which has thus been conceived as a complex adaptation with the function of communication. This is a conclusion that is supported by reverse engineering (Dennett 1995; Pinker 1997) and makes language the result of a complex project aimed at adaptation for communication and guided by natural selection.

Furthermore, the guiding idea that natural selection is the only process capable of explaining adaptive complexity (Bloom 1998) leads to a nativist conception of the human language.[2]

These kinds of conclusions awaken the critical spirit of the philosophers of science. Reverse engineering supports evolutionary explanations based on the

[2]The innatism of the linguistic components was claimed by Bickerton (Bickerton 1984) with regard to the transition from *Pidgin* to Hawaiian Creole and then confirmed by Pinker (Pinker 1994). Pinker himself rules out that chimpanzees can communicate through *American Sign Language* and are not capable of symbolic thinking (see Tartabini 2011: 33).

current function of a phenotypic trait: for example, it has been argued that the displacement of the larynx occurred as a function of the need to articulate phonemes (Lieberman 1988). However, is it still correct to support the role of selective implementation without considering the ecological conditions of the past that our ancestors found themselves facing? Is it plausible to invoke the concept of adaptation on the assumption of constant selective pressures along the time scale of the evolution of the genus *Homo* when organic selection phenomena (Baldwin 1896), which involve transformations of the environment and the construction of ecological niches (Laland et al. 2000; Odling-Smee et al. 2003), exist in nature? Is a naturalisation that involves a discontinuity between *Homo sapiens* and nature and that relies on the constant call of natural selection as the only explanatory cause of complexity the only viable path? Can language be spoken of in terms of a complex adaptation?

3 Darwin and Language

It is possible to seek an answer to these questions by retrieving what Charles Darwin wrote on language and trying to use the tools and conclusions that still remain valid today. Darwinian reflections on language focus on two main aspects:

(a) Evolutionary mechanisms of languages
(b) Evolution of the faculty of language[3]

3.1 Evolutionary Mechanisms of Languages: Languages Such as Species

Darwin went public with his views on the evolution of language in *Notebook 1836–1844* (Barret et al. 1987), *On the Origin of Species* (Darwin 1859, 1876), *The Descent of Man, and Selection in Relation to Sex* (Darwin 1874) and *The Expression of the Emotions in Man and Animals* (Darwin 1872). The discovery of some similarities shared by the cultural evolution of languages and the biological evolution of species led the author to metaphorically conceive languages as organisms, a theoretical position later inherited and developed by Charles Lyell (see Lyell 1863: 469–470).

Darwinian metaphors develop along these evolutionary trajectories:

– Analogies between the genealogical tree of species and human populations and the genealogical tree of languages (see Schleicher 1873: 13; Formigari 2012).

[3] By language faculty, we mean the physical and mental activities thanks to which it is possible to create languages and use them.

– Existence of a struggle for existence that involved languages and could cause their extinction or subsequent evolution.
– Components of words that could be analysed in terms of adaptations, analogies and homologies or vestigial aspects as in biological organisms.

Without focusing specifically on the origin of the first articulate languages (see Darwin 1859: 40, 310–311), the presence of an evolutionary tree allows the suggestion of the possibility of a classification of languages according to their genealogical origin or artificially based on other characters (see Darwin 1874: 90). Darwin opted for the first hypothesis, proposing a parallel between the genealogical arrangement of human varieties and the classification of spoken languages capable of delineating natural explanatory connections between extinct and current languages according to genealogical affinities (see Darwin 1859: 423–424; Darwin 1874: 148).

It also seemed possible to observe a struggle for survival between words and grammatical forms as an explanatory hypothesis applicable to the history of languages (see Darwin 1874: 91). This involved a selective perspective to explain the mechanisms of change of linguistic components, which resulted in the differential survival of dialects and languages, some of which have been preserved at the expense of others destined to expire (see Darwin 1874: 90–91): there is an ecological and social adaptation of languages which preserve them from extinction. Furthermore, the changes that led to the development of different languages took place gradually and progressively (see Barret et al. 1987: 581; Darwin 1874: 90).

Since vestigial traits of the original language remain in the following ones, homologies and rudiments are present in languages and, in addition to giving life to a series of comparisons between linguistic segments and vestigial aspects (see Darwin 1876: 402), were interpreted by the author as etymological evidence of historical-genealogical affinities (see Darwin 1874: 90).

Finally, between the species, there are analogies related to the phenomena of correlated growth and phonetic change (see Darwin 1874: 90).

3.1.1 Limits of the Darwinian Approach

Extending evolutionary analogies to characteristics common in languages and species is possible for some aspects related to the populational nature in the dynamics of transformation, such as geographic isolation, migration and drift with respect to parent stocks (see Pievani 2012: 162).

This still allows us to develop explanatory factors in an evolutionary key, such as isolation and migration, that influence and explain the variations found in the tree of language families. In the same way, it is possible to analyse the vestigial aspects that allow us to derive a historical-etymological ancestry. Furthermore, it is possible to evaluate the survival of a language, and of its speakers, in terms of adaptability, change and belonging to specific ecological-social interactions. When a language does not demonstrate these capacities for adaptation, it risks running into an

irreversible drowning in the oblivion of the internal symbolic universe it carries, a phenomenon corresponding to the extinction of a species.

It is precisely by observing these common points that it is possible to hold back the explanatory claims of the Darwinian metaphor (see Ellegård 1958: 291–292). In fact, if the replacement of a language with others in use or if the loss of the entire historical memory of a language is linked to the historical-social dynamics of a specific community, then Darwinian analogies in the biological field would attempt to understand phenomena linked mainly to social causes and describe events affected by cultural evolution.

A first critical observation, therefore, must address the ways of considering the variations that arise within a language. Unlike genetic variations, linguistic variations are not fixed; they do not guarantee the survival and reproduction of a language in relation to the ecological context of membership and development (Labov 2001).

Furthermore, the union of the Darwinian theory of variation with the theory of selection and their application to the evolution of languages give rise to another divisive point: natural selection does not pursue any progress, neither physical nor moral, but proceed by trial; therefore, the terms in which a selection of dialects and languages is conceived that mix and engage in a struggle for survival presuppose the nondirectional arising of variations. However, this last point above does not find its place in languages, the variations in which are often provided with a social orientation or in any case an intrinsic purpose or even political influence (Keller 1994).

There are also those who question the presence of a real selective process as a mechanism of language formation (see Mancini 2012: 135), underlining all the explanatory limits of the Darwinian genealogical approach. Additionally, we cannot forget the peculiarities of transformative dynamics in languages with respect to the Darwinian ecological dynamics between inheritance, speed of change, distribution in space (geographic and social) etc. (see Pievani 2012: 159).

Given the differences between words and biological entities, the capacities of adaptations of languages and the struggle for survival between words and grammatical forms are metaphors which cannot lay the foundations of an extension of the whole Darwinian model to the evolution of languages.

Darwinian evolutionary biological model does not properly take into account the fact that language is conditioned by formal and informal cultural institutions: for example, the proper and incorrect uses of terms, the meanings of words and the language changes arise as a result of the guidelines defined by royal academies of language and dictionaries. These kinds of cultural institutions can affect the propagation of a language change to achieve some social goal with their choice: in this way the introduction of a novel variant emerges as a teleological process, and the interplay between variation and selection is socially motivated.

Ultimately, the development of analogies and comparisons between the evolution of species and languages can be formulated at the cost of an extreme explanatory simplification of processes that concern, despite the use of common metaphors, substantially distinct contexts of cultural and biological evolution (Ramat 2012).

Cultural evolution is an autonomous process with respect to purely genetic dynamics or strictly adaptationist processes (Cavalli Sforza 2004; Cavalli Sforza

2010). In the history of culture, there is a differential spread of innovations involving a coevolution between genes and culture (Richerson and Boyd 2005): in this context we can find correspondences between the migratory waves of populations and the geographical displacement of genes and languages and more generally between phenomena of biological differentiation and cultural diversity. On the basis of these models, the Darwinian metaphor of the tree has been reformulated into a correlation between the tree of diversity of human populations and the tree of linguistic families (Cavalli Sforza et al. 1997) taking into consideration the presence of phenomena such as mutation, geographic isolation and genetic drift in both trees. But this reworking of analogies between the genetic field and the cultural domain, despite having an important heuristic value, highlights above all limits and differences: the dynamics relating to gene flow and hybridisation are distinct from the phenomena of semantic development and spread, from linguistic loan and creolisation. Furthermore, the speed of cultural changes differs from that one of genetic changes.

3.1.2 Philosophical Consequences

From the Darwinian point of view, the parallelism between the tree of the evolution of species and that of languages should be understood as an important methodological feature or scientific metaphor. The rigorous study of language cannot ignore evolutionary models capable of extending the explanatory power and falsifiability of hypotheses and excluding misleading philosophical convictions. However, in a kind of circularity, it can also be affirmed that the history of languages is valid as a model for studying the history of species, as the evolution of languages offered the nineteenth-century readers examples of gradual transformations and common descent. Specifically, it was possible to see homologies and vestigial aspects, thus fuelling the plausibility of Darwin's theory.

As will be explained further on, applying the Darwinian idea of the tree of life (see Sober 2000: 11) to the evolution of language, that is, existing languages trace backward in time to common ancestors, will prove a useful methodological tool for understanding the evolution of the faculty of language.

But there are other theoretical conclusions that are achieved thanks to the works of the British naturalist. The traditional classification of languages, in which essentialist presuppositions replace historical references, is abandoned. Instead, a genealogical criterion of classification is adopted for dialects and languages as well. This involves two orders of philosophical-methodological consequences: the first concerns the relationship between the complexity of a language and hierarchies of perfection, and the second reflects the relationship between the complexity of a language and the development of a society.

We can divide this overcoming of essentialism into two aspects: (1) complexity in terms of the perfection of a language and (2) complexity in terms of the cultural progress of a society.

Complexity in Terms of the Perfection of a Language

It is true that we can determine a differential level of complexity related to the domains of morphology, syntax, lexicon, phonology, pragmatics and semantics of a language and to various structural levels of each domain. However, is this sufficient to affirm the levels of primitiveness of a language with respect to another one?[4] This is a difficult methodological path to suggest (Ramat 2012). Actually, it is impossible to argue that primitive languages have remained in the last 6000 years of human history (Dahl 2004), and given the immeasurable distance between current human languages and the protolanguage of our ancestors, we are epistemologically incapable of establishing a parallel between the ontogenesis and phylogeny of languages (Givon 2009).

Thus far, the consequences of the Darwinian position remain confirmed. In evolution, the growth of internal complexity cannot be described in terms of value hierarchies because the structures of each variety are more or less functional in relation to the environmental context, whose ecology is changeable and contingent. If the selective pressures can change, then there are no perfect adaptations every time; rather, there are relatively optimal traits in a perennial relationship of compromise with other adaptations and systems belonging to the whole organism. For this reason, nature cannot be subdivided into inferior and superior species.

In this case, a legitimate extension of the Darwinian analogy leads us to affirm the same for languages (see Darwin 1874: 91). If it remains possible to recognise a relative complexity that is attributable to the individual structural domains of a language, such as syntax and phonetics, there is no absolute criterion that provides evidence to affirm that a language is primitive or less evolved than another more perfect one.

Complexity in Terms of the Cultural Progress of a Society

There is also another ancient belief that can be summarised in these terms: rudimentary and imperfect languages belong to human populations whose social organisation appears immature and primitive. Thus, a natural scale of progressive perfection was also created in the linguistic field, which sees, according to the considerations of Friedrich Schlegel, August Wilhelm Schlegel and most of the linguistics of the nineteenth century (see Ramat 2012: 87), the Indo-European languages as being at their apex.

As a consequence of this ideological equation, according to which decreased linguistic complexity means greater primitiveness from a cultural point of view, it

[4]For example, although he recognized the equal dignity of all languages and the genealogical classification, Wilhelm von Humboldt believed that comparison of different languages revealed the existence of hierarchy of complexity and different degrees of perfection: the Peruvian language was inferior to the Mexican one, while the Delaware language was presented as superior to the Burmese language (see Humboldt 2000: CI, 8, 16, 20–21).

was believed that the more finely articulated languages of primitive cultures should be found according to a creationist argument. That is, these languages are proof of a divine origin or a process of historical development that saw in languages the result of refined artistic elaboration by enlightened progenitors once belonging to current culturally and socially involuted societies.

Darwin reverses the reasoning by applying the naturalist perspective that does not classify the perfection of a species according to the degree of symmetry or complexity. Therefore, the regularity or complexity of a language is not proof of a superior culture or a divine origin, nor is the simplicity or rudimentary nature of a language proof that it can be traced back to a much more backward original culture that served as a receptacle for the formation of that same language (see Darwin 1874: 91–92).

After reading the writings of John Lubbock and Edward Burnett Tylor, Darwin was convinced that there is an anthropological parity to languages (Lubbock 1865; Tylor 1871), that it is not possible to establish a distinction between human races based on linguistic and/or mental hierarchies (see Darwin 1874: 179–180) and that the idea of a divine origin of language could not be supported (see Darwin 1874: 91–92).

Yet, typological thought adopted the idea that the complexity of language is a mirror image of the civilisation stage of a people to give impetus to the racist theses that seek a genetic correspondence to cultural diversities starting with the generic reworking of Darwinism in August Schleicher (Schleicher 1873) and in Chapter XXIII of Ernst Haeckel's *Natürliche Schöpfungsgeschichte* (Haeckel 1868).

It is still widely accepted today that there is no gene for language or grammar (Moro 2006), and although we must proceed with extreme caution in categorising languages as organisms subject to the laws of organic evolution, it is important to acknowledge that Darwin lived in a time when the claims regarding races were closely linked to the issue of slavery and propagation of the colonising languages. Furthermore, the merit of his having developed an anti-fixist historical approach that includes a space for a linguistic perspective based on an anti-racist key linked to the concept of common ancestry must also be acknowledged.

3.2 Evolution of the Faculty of Language

According to Darwin, language is half art and half instinct (see Darwin 1874: 610). Specifically, it is a particular case of art that, presenting aspects shared with animals and instinctive components in children, differs from all ordinary art. At the same time, language is a very special case of instinct because the existence of phases dedicated to the understanding and exercise of a language convinces us that it cannot be conceived as purely innate (see Darwin 1874: 87).

The hybrid nature of the faculty of language led the author of *The Descent of Man* to analyse its distinct components according to natural causes. He deals with the anatomical and physiological presuppositions, such as the brain and organs of

speech, but his reflection also dwells extensively on the elements of language understood as art, specifically all forms of vocal or mimic-gestural expression as components that have been learned by imitation and culturally transmitted.[5]

All these aspects can be traced back to the author's belief that language consists of a series of adaptations that developed gradually (see Barret et al. 1987: 599) as a result of individual variations devoid of direction that were maintained by the inheritance of advantages leading to the function of survival and reproduction.

Additionally, it is believed that these adaptations were thanks to the occurrence of other evolutionary engines, such as sexual selection, changes in functions,[6] dynamics of cultural evolution and implementations that can be explained through the principle of 'inherited effects of use' (see Darwin 1874: 87–88).

Rather than engaging in establishing the specific early languages, Darwin seemed to have wanted to reconstruct original physical and biological capacities combined with mental faculties superior to primates capable of anticipating and founding human words. The three most explored interpretative keys in the reflections of the author are continuity, imitation and coevolution.

3.2.1 Continuity

If non-human animals can communicate, if there are phenomena observable in the nature of communication between organisms, the articulation and understanding of articulated sounds and the connection of definite sounds and thoughts and if all higher mammals have anatomically endowed vocal organs similar to those of humans, then it would be increasingly difficult to support a strong discontinuist thesis. These ideas lay the foundations for a comparative study of human and animal faculties (see Darwin 1874: 84–86, 88–89; Barret et al. 1987: 558–559).

According to Darwin, semantics in animal languages do not consist of mechanical automatisms but are linked to the use of cognitive faculties such as memory (see Barret et al. 1987: 588), attention, understanding and cultural transmission that allow lower animals to learn an art (see Darwin 1874: 86; Barret et al. 1987: 264, 542, 569, 590–591). The same imitations of natural or animal sounds belong to the cognitive domain of some animals (see Barret et al. 1987: 533, 582–583) and monkeys (see Darwin 1874: 87; From Charles Darwin to Susan Darwin, 1 April 1838, Darwin Correspondence Project).

Current studies demonstrate that chimpanzees, just like sapiens, use a rich collection of postures, gestures and facial expressions and vocal calls to communicate emotions, intentions, warnings, alarms etc. (see Tattersall 2004: 57), and a

[5] Darwin had read *On the Origin of Language* (Wedgwood 1859), *Dictionary of English Etymology* (Wedgwood 1862) and chapters on language in *An Essay on the Origin of Language* (Farrar 1860): in these works he found the thesis that human language originated in the imitation of natural sounds.

[6] Gould and Vrba coined the term *exaptation* for results of an evolutionary change in function (Gould and Vrba 1982).

semantic decoding ability has been demonstrated in chimpanzees (Zuberbühler 2005). Not only primates do so: elephants, pinnipeds, bats and cetaceans are capable of vocal learning, an ability shared with several orders of songbirds, parrots and hummingbirds (Fishbein et al. 2019).

A phylogenetic continuity is considered possible between the imitative processes of the anthropomorph monkeys and the linguistic processes of *Homo sapiens* (see Di Vincenzo and Manzi 2012: 228). There is a multimodal communicative level where gestures and body movements are linked to words and form semantic units with a precise social functionality (see Aboitiz et al. 2020: 199–200). This level straddles the communicative processes of lower animals and the primitive stage of human language and allows the development of the communicative repertoire gradually through the intertwining of increasingly complex social and cultural dynamics.

As we will see later on, the absence of articulated language in chimpanzees should not be attributed to an unbridgeable gap in mental abilities (Duchin 1990).

3.2.2 Imitation

For the author, if primitive minds were considered devoid of complex conceptualisations, there remained, however, the possibility of associating expressions with mental states through perceptive and imitative bases.

The initial modality, which is the ancestor of language, would have been of an imitative nature and specifically the reproduction of cadences and musical cries through articulated sounds developed above all in moments of courtship. In this case, Darwin indirectly referred to the role of sexual selection and onomatopoeic imitation (see Barret et al. 1987: 568–569, 581, 593, 599; Darwin 1874: 86–87) but also to group selection in relation to alarm calls in monkeys (see Darwin 1874: 87; see Alter 2013:183).

Evolutionary dynamics also end up intertwining with each other; sexual selection would be one of the explanatory mechanisms for the development of the phonatory organs (see Darwin 1872:355; see Darwin 1874: 566–567), while the ‘changes in functions’ is an explanatory factor to which Darwin refers to conceive how any form of expression, which arose for independent functions, was then reused or co-opted for communication. This same reasoning was applied to the speech organs originally used for mating calls, and therefore for the reproduction of the species, and then converted to communication through articulated language and also implemented through the inherited effect of use (see Darwin 1872: 84–85, 355–356; Darwin 1874: 50).

Regarding the ability to emit musical sounds, although the author of *The Descent of Man* oscillated between the possibility that it is a practice inherited from our ancestors or a co-optation of the vocal organs previously adapted for a different function (see Darwin 1874: 571), he seemed much more determined to consider the hypothesis of a protolanguage. This protolanguage was used by the progenitors of our species before the formation of an articulated language corresponding to expressions capable of reproducing notes and rhythms or rudimentary melodies thanks to

musical instincts developed in the context of sexual selection (see Darwin 1874: 571–573; Barret et al. 1987: 595).

However, the author does not forget another fundamental element: besides the imitation and modification of the sounds of nature, the sounds of other animals and instinctive human cries, we must add facial expressions, gestures and signs to communicate with conspecifics (see Darwin 1874: 87; Barret et al. 1987: 573–574, 592).

Today, it is believed that prelinguistic forms of communication were used by *sapiens*, Neanderthals and other species and had developed in the dimension of sexual selection (Miller 2000) and maternal parental care (Falk 2009) thanks to the lengthening of the youthful period dedicated to social learning and imitation (see Pievani 2012: 164; Gärdenfors 2020: 264).

The author of the *Notebooks* had also underlined the importance of a 'pantomimic gesture' that was added to the intimate connection of sound and language (see Barret et al. 1987: 571). Authors such as Michael Corballis have studied the role of mimic-gestural communication in the evolutionary development of language to discover that a form of pantomime composed of ostensive gestures and referential expressions constituted the communicative circle of our Plio-Pleistocene ancestors (see Corballis 2014: 190).

Rediscovering the role of gestural communication is an important methodological outline[7]: the hypothesis that gestures are the precursors of linguistic communication and they are sufficient to create a context not at all devoid of grammatical norms, coherence and recursiveness came out from many studies on communication in chimpanzees. They are organisms capable of using gestures for requests for objects or actions, control of reactions and manipulation, and Michael Tomasello asserts that human cooperative communication arose in the intentional gestural communication of apes (see Tomasello 2009: 268). Additionally, the mirror system hypothesis also appears today as almost definitive support for the mimic-gestural origin of human language (Rizzolati and Arbib 1998; Rizzolati and Sinigaglia 2006; Di Vincenzo and Manzi 2012: 222–224).

As for the capabilities underlying protolinguistic forms of communication, Derek Bickerton suggests that they were shared by *Homo erectus* 1.6 million years ago (Bickerton 1995). On the other hand, the articulated language could not have been developed by Neanderthals but only with Cro-Magnon starting from 35,000 years ago (Lieberman 1991).

[7] In the second part of the *Essai sur l'origine des connoissances humaines*, Étienne Bonnot de Condillac maintains that it existed a language made up of natural gestures of the first men and women that also involved natural cries: the gestures became a collection of communications handed down and remembered. This would have been the basis for the development of the intellect and thinking that would have led to verbal language.

3.2.3 Coevolution

Finally, Darwin remained deeply convinced that the development of the faculty of language led to a reconfiguration of the mental and cognitive faculties of human beings, which over time would have determined a substantial difference between *sapiens* and other animal species (see Barret et al. 1987: 599; Darwin 1874: 610). The symbolic fixation of experience, implemented by the gesture-word combinations, through language would have acted in a circular and reciprocal manner on the physical and cultural preconditions of language's origin, implementing organs responsible for phonation (vocal organs, tongue, lips etc. See Darwin 1874: 9) and mental faculties also according to the theory of the inheritance of acquired characters (see Darwin 1874: 88–89).

Natural and cultural evolution would have determined a mutual restructuring of words and thoughts assuming there is no identity between thought and language, but that thought, to be fully articulated, needs the signs and support of language. In other words, language originates from an already developed mental activity but ends up retroactively amplifying the mental and cerebral capacities that in turn implement this faculty.

Although the neo-Darwinian synthesis rejects any reference to Lamarckism in regard to biological evolution, the Darwinian coevolutionary circle was confirmed and extended in a reciprocal feedback loop that involves cultural inheritance, the development of brain structures as a function of metabolic needs satisfied thanks to a renewed diet and favourable ecological conditions. All these factors are subject to the actions of natural selection, and exaptation would have allowed the transition from communication systems to systems of vocal articulation of language mediated by the fundamental role constituted by mirror neurons (see Di Vincenzo and Manzi 2012: 238).

3.2.4 Thresholds

It seems that animal vocalisations and systems of communication, even among non-human primates, constitute an embryonic form of evolution of human language (see Di Vincenzo and Manzi 2012: 219). The sophisticated alarm signals and vocal calls of monkeys offer a semantically impressive example of correlations between vocal expressions and the concept of predatory danger. These mechanisms are capable of guaranteeing a greater probability of survival to the communicative community that uses them (Krebs and Davies 2002; Arnold and Zuberbühler 2006). Yet, elements that seem to be found only in human communication are missing: recursive combinations of syntax. This poses an apparently insurmountable obstacle to the Darwinian thesis of the continuity between non-human primate communication and language.

If Darwin thought that syntax and semantics did not constitute thresholds, since the difference between animals and humans is the superior mental capacities of the

latter based on an almost infinite capacity for association among sounds and ideas (see Darwin 1874: 85–86), it seems today that syntax constitutes an insurmountable limit between the animal communication system and the language of the genus *Homo*.[8] However, there is still ongoing debate on recognising syntax as a cognitive module postulated by evolutionary psychology or whether it is an element linked to the historicity and sociality of the lexicon whose development cannot happen in the social and biological sphere, excluding any innate specialty.

Hauser (Hauser et al. 2002) considered natural selection to be incapable of explaining the generative and recursive computational dynamics of human syntax; it seems to be an exclusive element of the human cognitive domain.

How did it happen that a recursive system as refined and sophisticated as syntax evolved from an incipient form and not at all advantageous, given its primitive functionality, without being eliminated by natural selection? This constitutes Mivart's objection to Darwin as to how a sketch of a wing could be so advantageous as to be selected in the progenitors of birds or, to reformulate it according to the objections of the supporters of intelligent design, how 5% of an eye could be used so much so that it can be selected in our ancestors to become the complex system we have today.

Darwin's answer was the 'changes in functions' or 'co-optation' that we term today as *exaptation*. It is precisely from the co-optation of structures dedicated to the control of semantic and computational processes through structural learning by imitation that Di Vincenzo and Manzi claim language to have originated (see Di Vincenzo and Manzi 2012: 230).

3.2.5 Philosophical Consequences

Darwinian continuity helps to point out some important conclusions. Firstly, the overcoming of the philosophical tradition that maintained that the articulated word presupposed a conscious intention to communicate: the ontological gap between the human species and lower species was preserved in this conscious intention transformed into words.

The vocal forms of animals were rudimentary instinctive expressions with no words and thoughts to support them: human linguistic capacity was considered a dimension of human freedom that was historically exercised through the development of different and more articulated languages based on the processes of abstraction incommensurable with the primitiveness of animal verses (see Beer 1996: 111).

If animal noises remained irreducible to any linguistic qualification because they were purely irrational and linked to instinctive needs, the idea of a gradual evolution of language starting from elementary animal expressions could not be supported (see Müller 1861: 299).

[8] According to Stephen Anderson, a system of hierarchical, recursive syntactic combination has ever been found in animal communication (Anderson 2008).

According to Max Müller, the conceptual categories were exclusively within humans' reach thanks to language. The latter thus became proof of the impossibility of an evolution starting from the lower mental capacities of the monkey-like ancestors (see Alter 2013: 185).

Darwinian considerations on continuity have fuelled countless other studies, and today, the following conclusions have been reached: referential communication is also present in the animal world (Griffin 1976; Slobodchikoff 2002). Primates possess a vocal apparatus incapable of articulating words (see Tattersall 2004: 151–152), but there are no cognitive limits that prevent them from exhibiting semantic decoding skills and pragmatic abilities for communication (Tomasello and Call 1997; Zuberbühler 2005), and vocal learning is present in some groups of mammals and three groups of birds, specifically humans, bats, elephants, cetaceans, parrots, hummingbirds and songbirds (Jarvis 2004; Jarvis 2006).

If the communicative behaviour of certain animals is symbolic: the assumption of philosophers according to which animals are devoid of communicative intention is wrong, and the consequence is that there is no qualitative dichotomy but a profound quantitative difference in the complexity of the signal systems and within the range of intentions that divides the animal communication from human language (Griffin 1976). Moreover, there are consequences related to selection and coevolution.

As for selection, in the former philosophical vision, language was a specific feature; humanity was characterised by language, and it was not possible to conceive humanity without it.[9]

Within Darwinian theory, it is not necessary to conceive the faculty of language as something innate; language has led to significant evolutionary advantages in the natural history of mankind, but in the beginning, it may not have been an indispensable component to survival, unlike other far more important characteristics. This means that there may have been a stage in which humanity was humanity even without possessing the faculty for complex language. In other words, the original constitution of women and men and their ancestors is not reducible to the production of complex language because this idea depends on the evolution of anatomical preconditions and the consequent coevolution of the brain and language.

The Darwinian theory of selection and variation leaves the field open for the role of chance. There is no evidence of a directionality (Sproat 2011) or a project towards progress and refinement of complex language in Darwinian evolution, and this lays the groundwork for conceiving language as a natural faculty whose appearance

[9]We can find the idea that the faculty of language is what makes humanity unique in ancient philosophy, with Aristotle, till the modern age, with René Descartes, and although Jean-Jacques Rousseau proposed a study of the 'causes naturelles' of words, he conceived language as an element of distinction of humanity, excluding any continuity between animal language and human articulate expression (see Rousseau 1781: 5). But we also find it after Darwin's publications, for example, with Martin Heidegger who, in *Unterwegs zur Sprache*, states that humanity becomes humanity because of the innate and constitutive character of language (see Heidegger 1973: 27). In *Cartesian Linguistics* (Chomsky 1966), the author will take up this species-specific conception to found the intellectual organization obtained exclusively by humanity.

occurred according to random coordinates but at the same time necessarily linked to other biological phenomena, ecological constraints and adaptations at stake for the survival and reproduction of the species (see Renzi 2012: 193).

Regarding coevolution, the Darwinian concept of coevolution overcomes the belief of empiricism, according to which language, far from creating thought, is limited to expressing or reflecting thought (see Locke 1895: 3 footnote 2). However, above all, it resolves the critical point that sees the distancing between the two authors of the most controversial theory of the nineteenth century: Charles Darwin and Alfred Russel Wallace.

Wallace was sure that natural selection would be unable to explain the more complex mental and moral faculties of human beings (From Wallace to Darwin, 24 March 1869, Darwin Correspondence Project; see Wallace 1869: 391–394).

Although Darwin did not agree with Wallace (From Darwin to Wallace, 14 April 1869, Darwin Correspondence Project; From Darwin to Wallace, 26 June 1870, Darwin Correspondence Project), the problem remained: how can the increased mental capacities and brain volume of human beings be explained by resorting only to selection?[10]

The solution he reached was the mind-language coevolutionary connection: language can cause an increase in size and in the mental abilities of man, abilities that are then inherited and form a new starting point for the development of linguistic innovations again and again. In other words, language and mind develop this reciprocal and selective relationship: since they are exposed to different biological and cultural selective pressure, they change gradually, simultaneously or in a development of subsequent reactions, and they adapt to each other through the constant conservation of individuals presenting mutual and advantageous structure variations.

At present, the concept of coevolution according to which language adapts to the structures and capacities of the brain by causing a back-adaptive effect in the brain is supported by Michael Tomasello (Tomasello 1999) and Morten Christiansen and Nick Chater (Christiansen and Chater 2008): Dean Falk even thinks that the birth of music and language was concurrent and they affected the evolution of the two hemispheres of the brain for millions of years (see Falk 2015: 9).

[10]The answer to this question forced Darwin to face the skepticism of all philosophers who considered continuist positions as a reduction of human culture and moral to animal life. Johann Gottfried Herder had claimed with incisiveness that language as human invention is possible only when humanity is endowed with thought and rationality as species-specific qualities. Animals are devoid of reason and thinking, and they cannot lay the foundations of language: this philosophical attitude convinced philologists like Max Müller: 'Now I take my stand against Darwin on language, because language is the necessary condition of every other mental activity, religion not excluded, and I am able to prove that this indispensable condition of all mental growth is entirely absent in animals' (Letter from Max Müller to Duke of Argyll, 4 February 1875; see Müller 1902: 508–509).

4 Tree Thinking

The concept of common descent and the fact that the change in the lineages leading from the ancestor to the descendants is explainable on the basis of the idea of evolution are the heart of Darwinian concept of tree of life (see Sober 2000: 11).

According to the author of *The Origin of Species*, the relationships between ancestor species and descendant species form a single tree: studying the connections among present and past species means moving from the root to the tips of this tree.

The tree of life means that 'we may be all netted together' (see Barret et al. 1987: 229): all living beings, but also all the extinct forms which have ever existed, are connected along many lines of descent, or branches, converging to the common progenitors.

This is the pattern that evolution has produced and on the basis of which we can try to understand which species are descended from which others, to say when old characteristics originated and old ones disappeared and to trace the gradation in the state of the same organ and the gradations between the several species and groups of species in different closely connected orders (see Darwin 1877: 262). This is an important aspect of the tree thinking: trees can be used as methodological tools 'to present information on the distribution of traits among species' (see Baum et al. 2005: 980).

From a common ancestor descend a range of alternative results that are the genealogical related branches of the tree of life, and, as we said above, this same pattern is applicable to the evolution of languages, since the correct classification of the various languages now spoken throughout the world, which considers the various degrees of difference between the languages of the same stock and connects together all languages, extinct and recent, by the closest affinities, has to be strictly genealogical (see Darwin 1876: 371).

Therefore, Darwin's methodological path in classification, that is, '. . .characters being of real importance for classification only in so far as they reveal descent, we can clearly understand why analogical or adaptive characters, although of the utmost importance to the welfare of the being, are almost valueless to the systematist' (see Darwin 1859: 427), is effective to understand the evolution of language, as highlighted by Elliott Sober in these terms: 'To be sure, it is possible that each language independently evolved similar names for the numbers. But it is far more plausible to suppose that the similarity is due to the fact that the languages share a common ancestor. Once again, the reason this similarity is such strong evidence for a common ancestor is that names for given numbers are chosen arbitrarily. Arbitrary similarity, not adaptive similarity, provides powerful evidence of genealogical relationship' (see Sober 2000: 42).

Thinking about the tree of life, in other terms the 'tree thinking', is a feature of current evolutionary biology (see Sober 2000: 9), and it means that we have to study the phylogeny of each trait in full awareness of the fact that we need to consider nonadaptive explanations.

In other words: phylogenetic trees could be used as an efficient way to show historical relationships (see Baum et al. 2005: 979); therefore, tree thinking, that is, studying traits according the branches of phylogeny (see Pievani 2014: 140), helps to understand how comparative analysis on the basis of common descent can corroborate or confute an adaptive story which makes use of reverse engineering: in this way we can distinguish adaptation from simple inheritance of phylogeny.

Moreover, the processes and rhythms of evolutionary change can differ from one branch to the next of the tree of phylogeny of species, and this allows us to clarify the Darwinian pluralist explanatory approach.

According to Darwin when we are working out why a new trait evolved in a lineage, why old species became extinct or new ones came into existence, natural selection is the main process that explains what did occur in the tree of life, but it is not the only one: in the evolutionary descent, we can find sexual selection, group selection, changes in functions, the inherited effects of use and dynamics of cultural evolution and the influence of the 'mysterious laws of correlation of growth' (see Darwin 1859: 143), but today we can also consider the role of genetic drift, ecological changes and other explicative views. In one branch of the tree of life, a trait of a species has evolved under selective pressures, but in another branch of the tree, that is, in the evolutionary history of another species, the same character went through a shift of function or other mechanisms (see Pievani 2014: 139): adaptationist explanation is one among the different explicative alternatives Darwin proposed.

We have to keep this conception in mind when we are understanding language as a mosaic of different cognitive, inferential and physical abilities with different evolutionary, and cultural evolutionary, histories (Planer and Godfrey-Smith 2021).

5 The Consilience Approach

Thus far, the independent mechanisms that we have seen to compose the faculty of language have been observed in continuity with the animal world. However, syntax has posed a problem that can give rise to a contrast between the continuist and the innatist positions. The problem is trying to understand how iteration, sequency and recursiveness primitively present in the simple combinatory system of animal communication could have evolved into syntax. Postulating a module of syntax, as is proposed by evolutionary psychology, leads, as we saw at the beginning, to a discontinuist result.

Yet, according to Tattersall, a comparison with the primate brain allows us to affirm that there are no unique brain structures which support the thesis of the superiority of our cognitive abilities (see Tattersall 2004: 68). Moreover, evidences for compositional syntax have been found in birds and primates (Fishbein et al. 2019).

However, a comparison with Tecumseh Fitch's consilience approach legitimises an epistemological extension of naturalistic explanation regarding the origin of

language, making it open to the assimilation of Darwinian tree thinking. This allows us to overcome difficulties generated by this kind of selective gradualism (Fitch 2011; Fitch 2012).

According to the author, the faculty of language is not conceivable as a single and complex organ; rather, he conceives language as a multi-component system to which it is possible to apply the Darwinian multimodal approach. The mechanisms that led to the origin of this faculty have had an independent evolutionary history that has led them to interact and integrate. Furthermore, some of these have homologies or analogies in other animal species, others have been extraordinarily implemented during cultural evolution, others appear as adaptations for natural selection, others 'reflect deep developmental and phylogenetic constraints on the physiological and neural mechanisms underlying linguistic behaviour' (see Fitch 2012: 623) and still others are the result of exaptation. None of these, taken alone, offer the access key to understanding and defining language, but despite such a pluralistic approach, the development of these components can be analysed in terms of phylogeny.

Fitch formulates three hypotheses for the evolution of language and finds an access key to the mystery of syntax by beginning with Darwinian exaptation that he reformulates into these terms: 'Complex structures, evolved in one functional context, can change their function, and be put to work in new domains, often carrying with them traces or constraints due to their prior function' (Fitch 2011).

This may have happened for what has characterised human speech abilities, but is missing in chimpanzees, such as laryngeal control. That is, the direct cortical control over the larynx is the result of exaptation of the cortico-spinal tract in primates (see Fitch 2011: 5); but there is also the brain circuit provided by the arcuate fasciculus, which connects syntactic and lexical regions: previously evolved for vocal imitation, now it makes syntactic comprehension and semantic interpretation possible (see Fitch 2011: 6); finally the author considers that the evolution of specific cortical regions involved in syntax, specifically the specialised components of Broca's area, finds precursors in non-linguistic regions appointed to motor control and motor planning and/or in multimodal areas involved in visually guided aspects of gaze when it is exposed to social pressures (see Fitch 2011: 7).

In this way, he clarifies how exaptation orients our way of conceiving the Darwinian evolution of complex structures such as language:

- There is no preordained and infused direction in evolution.
- Evolution works starting from materials made available by nature.
- The change of function presents a continuist Darwinian theory regarding evolutionary change, but it presents also a 'discontinuist Darwin' regarding the function of adaptations. Although modern evolutionary synthesis has weakened Darwinian gradualism, with exaptation, we can no longer think in terms of an organ made or built once and for all to perform a certain function.

6 Conclusion

Starting from the central idea that the selectionism of evolutionary psychology undervalues Darwinian tree thinking, we have found that current explanations, inspired by Darwinian intuitions, explore all the explanatory possibilities regarding the origin of human faculty of language. The modules view, according to which human behaviour and cognitive systems are defined by a number of innate, separate and distinct cognitive modules, originates from the philosophical perspectives of Chomsky (1980) and Fodor (1983), and it is inherited and developed in the evolutionary psychology approaches of Tooby and Cosmides (1989) and Pinker (1997). In opposition to this view, naturalising means accepting that the human mind shares the uniqueness and variety that evolution has endowed with all living species. It is true that apes show only the starting point of the cognitive skills of our ancestors, but in the research on the evolutionary origin of human cognitive abilities, the comparative study with the mental abilities of apes remains very important (see Tattersall 2004: 72).

The comparison with Fitch's position has allowed us to understand how enlightening the phylogenetic reconstruction can be when it is in accordance with Darwinian explanatory pluralism: tree thinking, that is, an analysis of traits based on the study of genealogical relationships and comparative method (see Pievani 2014: 139), allows us to overcome the difficulties of adaptationism of evolutionary psychology but forces us to a philosophical reformulation that renounces the innatist and discontinuist theses in the history of philosophy.

Starting with Darwinian works, this article attempted to establish the possible preconditions of the faculty of language both in phylogenetic and ontogenetic terms. Intertwining Darwinian hypotheses with current conclusions made it possible to identify the most plausible evolutionary models. From these, this work drew methodological conclusions regarding the nature of language and *Homo sapiens*.

The new epistemological model requires the abandonment of abstract speculations regarding the transcendental conditions of linguistic phenomena, which must instead be traced back to the description, understanding and interpretation of historical facts and empirical and experimental content conditioned by evolutionary processes.

Furthermore, applying the Darwinian theses, it is possible to observe that different lines of research have developed from them to reach multiple outcomes. Although other non-Darwinian lines of research have proven to be important in clarifying the implications of a naturalisation, Darwinian pluralism remains a further configuration of linguistic naturalism from which philosophy can benefit to set up a methodological programme that touches the following points:

(a) Language is a mosaic of cognitive abilities and physical structures connected in a sophisticated way but conceivable and explainable in terms of a multimodal naturalistic approach based on the recourse to Darwinian evolution.
(b) Naturalisation does not consist in a study aimed at disclosing the origin of language, but it is a research for the psycho-physical and cultural conditions

and the ecological and social premises needed for the rising of the faculty of language and its development in current practices. Language remains the result of rules inscribed in the biological organisation of our body during the evolution of species. Philosophy suspends judgement about an extra-natural, metaphysical, theological or transcendental foundation, and philosophers find themselves consolidating physical-anatomical, chemical, biological, historical knowledge etc., because these skills constitute the irreducible elements without which origins of language cannot be described.

(c) The principles of language are not the starting points of our 'being human', but points of arrival of a pluralism of evolutionary factors common to all animals.

(d) Speaking of a species-specific characteristic does not mean understanding language according to an ontological gap. The use of the comparative approach forces philosophers not to consider human specialisations by attributing them a special or unique meaning. Confusing the research into human evolutionary history with the search for *essence* or characteristics that make the human *form* or soul unique is methodologically wrong. Naturalising means welcoming what unites us with other animals and with nature in general but at the same time discovering what are the most properly human specialisations through the perspectives of biological evolution and cultural evolution.

(e) Furthermore, the transformations underlying the formation of articulated language based not on the implementation of the communicative function but in terms of the evolutionary domains of defence from predators and dangers, nutrition and sexuality must be considered.

(f) The complex activities of language also depend on the multiple reuse or functional co-optation of traits with different functions over time, and this has two important consequences:

- The components of language can develop on the basis of two distinct dimensions: some are characterised by a continuity of function and others by exaptation or by an evolutionary process starting from non-communicative precursors.
- The shift of the current utility of a characteristic from its historical origin is a methodological conclusion of the concept of exaptation, which helps to solve the critical question generated by the adaptationist paradigm of evolutionary psychology and the use of reverse engineering.

(g) If it is possible to attribute to animal species the mental abilities necessary for the development of language. It is also true that we can't find in animal kingdom any developed form of language even remotely closed to that one of the *sapiens* species. This opens a cultural and social dimension for naturalisation. Cultural evolution and its dynamics are grafted into the evolutionary path of language after the body structures have reached levels of energetic, cognitive and anatomical sustainability capable of giving rise to the use of sounds as a function of symbolic thought. The encounter between the biological and cultural dimensions has launched a coevolutionary circle responsible for the enhancement of many aspects of language. Social and cultural processes of evolution are different from

biological processes, but they're not separate from them. Biological evolution is closely integrated into the sociocultural dimension of the *sapiens*, and this research domain finds remarkable insights into the neurosciences: with mirror neurons hypothesis, for example, a biological-cultural dimension opens up where the brain not only resorts to symbolic thought but also establishes an interactive understanding on the basis of a motor process as a fundamental tool to enhance learning and the social and pragmatic development of language.

References

Aboitiz F, Osorio S, Henríquez-CH RA (2020) A neural code for multimodal language processing and its origins. In: Ferretti F, Adornetti I (eds) Paradigmi: the speaking brain, the origin of language from a cognitive standpoint. Il Mulino, Milano

Alter SG (2013) Darwin and language. In: Ruse M (ed) The Cambridge encyclopaedia of Darwin and evolutionary thought. Cambridge University Press, Cambridge, pp 182–187

Anderson SR (2008) Doctor Dolittle's delusion: animals and the uniqueness of human language. Yale University Press, New Haven

Arnold K, Zuberbühler K (2006) The alarm-calling system of adult male putty-nosed monkeys. Cercopithecus nictitans martini Animal Behaviour 72(3):643–653

Baldwin JM (1896) A new factor in evolution. Am Nat 30(354):441–451

Barret PH, Gautrey PJ, Herbert S, Kohn D, Smith S (eds) (1987) Charles Darwin's notebooks, 1836–1844, geology, transmutation of species, metaphysical enquiries. Cornell University Press, New York

Baum DA, De Witt SS, Donovan S (2005) The tree-thinking challenge. Science 310(5750): 979–980

Beer G (1996) Open fields: science in cultural encounter. Oxford University Press, Oxford

Bickerton D (1984) The language bioprogram hypothesis. Behav Brain Sci 7(2):173–188

Bickerton D (1995) Language and human behavior. University of Washington Press, Seattle

Bloom P (1998) Some issues in the evolution of language and thought. In: Cummins D, Allen C (eds) The evolution of mind. Oxford University Press, Oxford, pp 204–223

Cavalli Sforza LL (2004) L'evoluzione della cultura. Codice Edizioni, Torino

Cavalli Sforza LL (2010) La specie prepotente. Editrice San Raffaele, Milano

Cavalli Sforza LL, Menozzi P, Piazza A (1997) Storia e geografia dei geni umani. Adelphi, Milano

Chomsky N (1966) Cartesian linguistics. A chapter in the history of rationalist thought. Harper and Row, New York

Chomsky N (1980) Rules and representations. Columbia University Press, New York

Christiansen M, Chater N (2008) Language as shaped by the brain. Behav Brain Sci 31(5):489–509

Corballis M (2014) The word according to Adam: the role of gesture in language evolution. In: Seyfeddinipur M, Gullberg M (eds) From gesture in conversation to visible action as utterance: essays in Honor of Adam Kendon. John Benjamin Publishing Company, Amsterdam, pp 177–197

Dahl Ö (2004) The growth and maintenance of linguistic complexity. John Benjamins Publishing Company, Amsterdam/Philadelphia

Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, 1st edn. John Murray, London

Darwin C (1872) The expression of the emotions in man and animals, 1st edn. John Murray, London

Darwin C (1874) The descent of man, and selection in relation to sex, 2nd edn. John Murray, London

Darwin C (1876) The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, 6th edition with additions and corrections. John Murray, London
Darwin C (1877) The various contrivances by which orchids are fertilised by insects, 2nd edn. John Murray, London
Dennett D (1995) Darwin's dangerous idea: evolution and the meanings of life. Simon and Schuster, New York
Di Vincenzo F, Manzi G (2012) L'origine darwiniana del linguaggio. In: Banfi E (ed) Sull'origine del linguaggio e delle lingue storico-naturali, un confronto tra linguisti e non linguisti, Atti del primo convegno interannuale di studi della società di linguistica italiana (SLI), vol 2013. Bulzoni Editore, Roma, pp 217–247
Duchin L (1990) The evolution of articulate speech: comparative anatomy of the oral cavity. J Hum Evol 19(6–7):687–697
Ellegård A (1958) Darwin and the general reader. The reception of Darwin's theory of evolution in the British periodical press, 1859–1872. Almqvist & Wiksell, Göteborg
Falk D (2009) Finding our tongues. Mothers, infants & the origins of language
Falk D (2015) Lingua madre Cure materne e origini del linguaggio (Trans: Dossena P). Bollati Boringhieri, Torino. Italian Edition
Farrar FW (1860) An essay on the origing of language. Murray, London
Fishbein AR, Fritz JB, Idsardi WJ, Wilkinson GS (2019) What can animal communication teach us about human language? Phil Trans R Soc B 375(1789):20190042
Fitch T (2011) The evolution of syntax: an Exaptionist perspective. Front Evol Neurosci 3(9):1–12
Fitch T (2012) Evolutionary developmental biology and human language evolution: constraints on adaptation. Evol Biol 39:613–637
Fodor J (1983) The modularity of mind. An essay on faculty psychology. MIT Press, Cambridge (MA)
Formigari L (2012) L'origine del linguaggio. Ricognizioni storiche e valenze epistemologiche. In: Banfi E (ed) Sull'origine del linguaggio e delle lingue storico-naturali, un confronto tra linguisti e non linguisti, Atti del primo convegno interannuale di studi della società di linguistica italiana (SLI), vol 2013. Bulzoni Editore, Roma, pp 13–22
Gärdenfors P (2020) From pantomime to protolanguage. In: Ferretti I, Adornetti I (eds) Paradigmi: the speaking brain, the origin of language from a cognitive standpoint. Il Mulino, Milano
Givon T (2009) The genesis of syntactic complexity. John Benjamins Publishing Company, Amsterdam/Philadelphia
Gould SJ, Vrba ES (1982) Exaptation-a missing term in the science of form. Paleobiology 8(1): 4–15
Griffin DR (1976) The question of animal awareness. The Rockefeller University Press, New York
Haeckel E (1868) Natürliche Schöpfungsgeschichte. Gemeinverständliche wissenschaftliche Vorträge über die Entwicklungslehre im Allgemein und dieje-nige von Darwin, Goethe, und Lamarck in besonderem. Berlag von Georg Reimer, Berlin
Hauser MD, Chomsky N, Fitch T (2002) The Faculty of Language: what is it, who has it, and how did it evolve? Science 298(5598):1569–1579
Heidegger M (1973) In cammino verso il Linguaggio (Trans: Caracciolo A). Mursia, Milano. (Italian Edition)
Humboldt W (2000) La diversità delle lingue (Trans: Di Cesare D). Laterza, Lecce. (Italian Edition)
Jarvis ED (2004) Learned birdsong and the neurobiology of human language. Behavioral Neurobiology of Birdsong 1016(1):749–777
Jarvis ED (2006) Selection for and against vocal learning in birds and mammals. Ornithol Sci 5(1): 5–14
Keller R (1994) On language change. The invisible hand in language. Routledge, London-New York
Krebs J, Davies N (2002) An introduction to behavioral ecology. Blackwell Science Ltd, Oxford
Labov W (2001) Principles of linguistic change, 2, social factors. Blackwell, Malden–Oxford

Laland K, Odling-Smee FJ, Feldman MW (2000) Niche construction, biological evolution and cultural change. Behav Brain Sci 23(1):131–175
Lieberman P (1988) On human speech, syntax and language. Hum Evol 3:3–18
Lieberman P (1991) Uniquely human: the evolution of speech, thought, and selfless behavior. Harvard University Press, Cambridge
Locke J (1895) An essay concerning human understanding, 2 vol. Clarendon Press, Oxford
Lubbock J (1865) Prehistoric times. Williams and Norgate, London
Lyell C (1863) The geological evidences of the antiquity of man. John Murray, London
Mancini M (2012) Il paradosso darwiniano: convergenze e divergenze di paradigma. In: Banfi E (ed) Sull'origine del linguaggio e delle lingue storico-naturali, un confronto tra linguisti e non linguisti, Atti del primo convegno interannuale di studi della società di linguistica italiana (SLI), vol 2013. Bulzoni Editore, Roma, pp 105–142
Miller G (2000) The mating mind: how sexual choice shaped the evolution of human nature. Doubleday, New York
Moro A (2006) I confini di Babele. Il cervello e il mistero delle lingue impossibili. Longanesi, Milano
Müller G (1902) The life and letters of the right honourable Friedrich Max Muller, 2 vols. Longmans, Green, London
Müller M (1861) Lectures on the science of language delivered at the Royal Institution of Great Britain in April, May, and June, 1861, 2nd edn. Charles Scribner, London
Odling-Smee FJ, Laland K, Feldman M (2003) Niche construction. The neglected process in evolution. Princeton University Press, Princeton
Pievani D (2012) L'evoluzione del linguaggio e "la grande espansione umana": nuovi intrecci fra genetica e linguistica. In: Banfi E (ed) Sull'origine del linguaggio e delle lingue storico-naturali, un confronto tra linguisti e non linguisti, Atti del primo convegno interannuale di studi della società di linguistica italiana (SLI), vol 2013. Bulzoni Editore, Roma, pp 153–168
Pievani D (2014) Evoluti e abbandonati. Sesso, politica, morale: Darwin spiega proprio tutto? Giulio Einaudi Editore, Torino
Pinker S (1994) The language instinct. W. Morrow and co., New York
Pinker S (1997) How the mind works. W. W. Norton and Company, New York
Pinker S, Bloom P (1990) Natural language and natural selection. Behav Brain Sci 13(4):707–727
Planer RJ, Godfrey-Smith P (2021) Communication and representation understood as sender-receiver coordination. Mind Lang 36:750–770
Ramat P (2012) Are all languages equally complex? In: Banfi E (ed) Sull'origine del linguaggio e delle lingue storico-naturali, un confronto tra linguisti e non linguisti, Atti del primo convegno interannuale di studi della società di linguistica italiana (SLI), vol 2013. Bulzoni Editore, Roma, pp 87–104
Renzi L (2012) Cambiamento linguistico e teoria dell'evoluzione della specie. In: Banfi E (ed) Sull'origine del linguaggio e delle lingue storico-naturali, un confronto tra linguisti e non linguisti, Atti del primo convegno interannuale di studi della società di linguistica italiana (SLI), vol 2013. Bulzoni Editore, Roma, pp 185–195
Richerson P, Boyd R (2005) Not by genes alone: how culture transformed human evolution. The University of Chicago Press, Chicago
Rizzolati G, Arbib MA (1998) Language within our grasp. Trends Neurosci 21:188–194
Rizzolati G, Sinigaglia C (2006) Sei quel che Fai, Il Cervello che Agisce e i Neuroni Specchio. Raffaello Cortina, Milano
Rousseau JJ (1781) Essai sur l'origine des langues, édition A. Belin, Paris
Schleicher A (1873) Die Darwinsche Theorie und die Sprachwissenschaft, Offenes Sendschreiben an Herrn Dr. Ernst Häckel, o. Professor der Zoologie und Director des Zoologischen Museums an der Universität Jena. Zweite Auflage, Hermann Böhlau, Weimar
Slobodchikoff C (2002) Cognition and communication in prairie dogs. In: Bekoff M, Allen C, Burghardt G (eds) The cognitive animal: empirical and theoretical perspectives on animal cognition. MIT Press, Cambridge, pp 257–264

Sober E (2000) Philosophy of biology. Westview Press, Boulder, Colorado
Sproat R (2011) Phonemic diversity and the out-of-Africa theory. Linguist Typology 15(2): 199–206
Tartabini A (2011) Sull'origine del linguaggio. In: Bianchi N (ed) Dialoghi sulle lingue e sul linguaggio. Pàtron Editore, Bologna
Tattersall I (2004) Il Cammino dell'Uomo (Trans: Comoglio M). Garzanti Editore, Milano. (Italian Edition)
Tomasello M (1999) The cultural origins of human cognition. Harvard University Press, Cambridge (MA)
Tomasello M (2009) Le origini della comunicazione umana (Trans: Romano S). Raffaello Cortina Editore, Milano. (Italian Edition)
Tomasello M, Call J (1997) Primate Cognition. Oxford University Press, Oxford
Tooby J, Cosmides L (1989) Evolutionary psychology and the generation of culture. Ethol Sociobiol 10:29–49
Tylor EB (1871) Primitive culture. Researches into the development of mythology, philosophy, religion, language, art and custom, John Murray, London
Wallace A (1869) Review of principles of geology by Charles Lyell. Q Rev 126:359–394
Wedgwood H (1859) On the origin of language. Trübner & Co., London
Wedgwood H (1862) Dictionary of English etymology. Trübner & Co., London
Zuberbühler K (2005) The phylogenetic roots of language: evidence from primate communication and cognition. Curr Dir Psychol Sci 14(3):126–130

Part III
Gene and Genotype Metaphysics

A New Perspective on Type-Token Distinction in the Genotype and Phenotype Concepts

David Ricote and Ignacio Maeso

Abstract The genotype-phenotype distinction is a core theoretical framework in biology. Genotype and phenotype classify biological entities or organisms as groups of individuals (the tokens) that share a common identity (the type), but so far it has not been investigated how these two concepts classify their tokens as types. However, this is key to defining which biological entities can and cannot be denoted by genotype and phenotype, a prerequisite to properly understand the so-called nongenetic inheritance. Here, we analyze type-token relations in genotype and phenotype and propose a new framework to differentiate these concepts by how they classify their tokens. We first argue that genotypes should be defined independently of genes, distinguishing genotype (classification of whole inherited structures) from genetype (classification of genes). Second, we propose that genotypes are natural kinds because they replicate and conserve their type-identity by self-templating, independently from being intentionally classified by humans. Conversely, phenotypes would not constitute natural kinds, as their tokens are intentionally classified. Finally, the identification of self-templating as a fundamental genotypic property opens new avenues to include additional inherited structures that are different to the genome, as parts of the full genetic identity of organisms, of their genotype.

D. Ricote (✉)
Andalusian Centre for Developmental Biology (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain

Departamento de Genética, Fisiología y Microbiología, Universidad Complutense de Madrid (UCM), Madrid, Spain
e-mail: dricote@ucm.es

I. Maeso (✉)
Andalusian Centre for Developmental Biology (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain

Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, Spain

Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona, Spain
e-mail: imaeso@ub.edu

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_11

Keywords Type-token · Genotype · Phenotype · Genetype · Genotoken · Phenotoken · Nongenetic inheritance · Membranome

1 Introduction

Genotype and phenotype are two core concepts in biology (Sapp 1983; Alberch 1991). They are profusely used in biology and biochemistry manuals, where they are illustrated with abundant examples. Despite their importance, and in contrast to the related concept of gene, both terms have generally received little philosophical attention (Taylor 2018); in particular, an in-depth discussion on how they are characterized *as concepts* is still missing. This characterization is essential to avoid inconsistencies when using these two concepts and to clearly define which material entities or instances can and cannot be denoted by them. This is a crucial biological problem because what we consider as genotype also defines what we can consider as genetic. In recent years, an increasing number of non-genomic hereditary systems are being acknowledged, such as certain cell membranes, histone modifications, cytoskeletal organizations, or learned behaviors (Jablonka and Lamb 2005; Bonduriansky and Day 2020). However, these hereditary factors are generally considered as nongenetic and therefore excluded from the genotype-phenotype framework, which constitutes an obstacle to properly understanding these additional inheritance systems and their biological relevance. There are some exceptions, with authors such as Cavalier-Smith, who coined the terms *genetic membranes* and *membranomes* to refer to the hereditary nature of certain cell membranes: the cytoplasmic and mitochondrial membranes must be inherited from a mother cell, in contrast to other membranes, such as those of vacuoles, that can be synthesized de novo (Cavalier-Smith 2004). The use of the term *genetic* to qualify these inherited cell membranes would imply that, in addition to genomes, there are other structures that can also be an integral part of the genetic makeup of living beings, part of their genotypes. But, can we really incorporate non-genomic hereditary structures into the genotype-phenotype framework? To assess these problems, we propose here a review and reframing of both genotype and phenotype, paying special attention to the relations of identity implicitly assumed in most genotype and phenotype allusions. These implied identity relations can be clearly seen in this concise and well-known definition of both concepts by Lewontin (1992, p. 137) (for an extended discussion on genotype and phenotype distinction, see Taylor and Lewontin (2017)):

> The "phenotype" of an organism is the class of which it is member based upon the observable physical qualities of the organism, including morphology, physiology, and behaviour at all levels of description. The "genotype" of an organism is the class of which it is member based upon the postulated state of its internal hereditary factors.

Here, Lewontin is characterizing genotype and phenotype as classes used to categorize particular material instances: *observable physical qualities* makes reference to traits or biological functions like skin pigment or metabolism processes. Those qualities must be perceivable characteristics of living beings. By contrast the

genotypes as *internal hereditary factors* refer to the specific configuration of the inheritable systems contained in the organism which will determine the qualities of phenotypes. The most paradigmatic example of genotype is the genome sequence present in DNA macromolecules. For referring to these particular macromolecules as instances, a few authors have sometimes used the term *token* (Lewontin 1992, p. 139):

> Individuals belonging to those classes are tokens of those types. The actual physical set of inherited genes, both in the nucleus and in various cytoplasmic particles such as mitochondria and chloroplasts, make up the genome of an individual, and it is the description of this genome that determines the genotype of which the individual is a token.

There Lewontin used type-token distinction to differentiate the relation of identity (genotype) from the genome sequence being identified (token or *genotoken*). Nevertheless, how this relation of identity works as a concept is taken for granted, as it is mentioned without any further discussion. In fact, this distinction between tokens and types is very rarely mentioned in scientific literature, and there is currently a conceptual problem where genotype and phenotype are used indistinctly to refer to both tokens and types. And perhaps more importantly, it is overlooked if genotype and phenotype concepts classify their tokens in the same manner.

In this chapter, we investigate this relatively neglected aspect of the genotype-phenotype distinction, namely, how the type-token relations are formed in each of these two concepts. The genotype will be the center of our attention, since we hypothesize that the genotype concept classifies tokens as natural kinds. To develop this hypothesis, we first define which tokens are of the genotype concept. Under our perspective, the genotype concept classifies entire inherited systems (i.e., whole genomes) rather than the arbitrarily defined segments (i.e., the genes) that are part of these hereditary systems. Second, we propose that the essential property of these whole inherited systems is that they reproduce through generations by guiding or self-templating their own replication. Therefore, the genotype concept classifies individuals that are naturally identified as types. This self-template constitutes a material and genealogical link between the reproduced tokens, and therefore their replicated identity is independent of being classified by external agents such as humans. Crucially, this applies not only to genomes but also to other inherited structures, such as genetic cell membranes, centrosomes, or chromatin modifications. Finally, we argue that in contrast to genotype, phenotype classifies and individuates tokens depending on the external cognition and pragmatic purposes of humans. Since these purposes may vary according to the requirements and praxes of different biological disciplines, to define traits and biological phenomena as phenotokens and their classification into phenotypes is a much more complex and context-dependent endeavor.

Thus, we propose that our clarification of the type-token relation in the genotype concept could be the basis for a new way of conceiving genotypes as well as their ontological distinction from phenotypes. We propose that our new perspective about genotype and phenotype concepts has deep implications on how we define

hereditary systems and how different forms of inheritance other than genomes could be incorporated into the genotype concept.

2 Genotype and Phenotype Are Type Concepts

2.1 Type-Token Distinction and Natural Kinds

We first need to briefly introduce the type-token distinction before discussing it within the context of genotype and phenotype. Type-token difference is a perspective of the classic philosophical debate about the relation between concepts or categories (types) and their particular material representations or instances (tokens). This distinction is oriented to the processes of classification of individuals by following different rules that can be potentially adapted to all kinds of ontologies. This is especially insightful for disciplines like semiotics or biological sciences because they deal with myriads of particulars to be classified and identified, like spontaneously uttered words or new specimens from unexplored environments (see Box 1 for different kinds of type-token distinctions).

Type-token distinction is made clear with a classical example from linguistics. In the phrase, *A rose is a rose*, there are three word types (a, rose, is) and five token words (two *a*, two *rose*, and one *is*). By emphasizing this distinction, the difference between the reference of a material entity as a token and the reference of its relation of identity with other tokens of the same type is made explicit. As a result, in the previous example, the token-word *rose* is the particular material entity where the token is present, such as a particular configuration of LEDs displaying this word on a screen or the ink in a printed version of this word (as just two examples of the multiplicity of possible material entities). By contrast, the type-word *rose* refers to the classification of the different token-words *rose* within a single category shared by literate humans. This identification made by literate humans is also connected with other type words, such as the categories of flower nouns or the general category of nouns in English. Making explicit this process of identifying successive tokens under one type is a main feature of the type-token distinction. It points out the difference between the reference to the tokens-instances and the reference to their classification as types.

Such distinction paves the way to question whether the tokens are classified by themselves or, on the contrary, they are classified in a nominalist fashion, which is connected with the old question about *natural kinds* – in other words, to question who are the subjects or entities (agents) that exert this classification. We follow the simplest definition of natural kind, a classification that "corresponds to a grouping that reflects the structure of the natural world rather than the interests and actions of human beings" (Bird and Tobin 2022). A perfect example of nominalism is the aforementioned example of the word *rose*, where literate humans are the agents identifying a discrete spatial structure as a word token, forming thus a type that connects the word-token with other similar words *rose* displayed in different places.

This process of identification is exerted always by literate humans, which also has the potential capacity of identifying other word-tokens with a very different structure, such a handwritten word-token. Therefore, it is clear that the reference of word-type depends on human literacy. By contrast, natural kinds, as opposed to nominal kinds, imply descriptions of tokens that are organized independently from human cognition in discrete types, like the elements in physics and chemistry or the species in biology. The central question is as follows: Is it possible that other living beings different from humans could generate their own type-relations? In biological disciplines, where type concepts are widely used, we can find a neat example of this duality through two different type concepts: type specimen and type species. The main difference between these two type concepts is that type specimen makes reference to an explicitly artificial classification by using a carefully chosen token as a representation of all characteristics of the type, e.g., holotypes, paratypes, etc. in museum collections, being the type specimen a type and a token at the same time. By contrast, type species refers to the result of the reproduction of biological individuals (tokens) that dynamically maintain their type through generations, independently of being classified by biologists. Accordingly, type species have been proposed many times as natural kinds, although this is a matter of intense discussion (Hull 1976, 1978; Ereshefsky 1998; Ellis 2001). By contrast, type specimens are admittedly types made up by biologists to show phenotypic characteristics of a species; therefore, type specimens cannot be natural kinds.

2.2 *Genotype and Phenotype Concepts under the Type-Token Distinction*

As we stated before, genotype and phenotype are not generally framed under the type-token distinction. In biology textbooks and other generalist sources, we find at least two senses of genotype and phenotype: *(i)* the material instances, such as the molecular sequence of the genome in the case of genotype and any trait or function such as the presence of particular pigments in the skin and hair in the case of phenotype, and *(ii)* the classification of these material instances, for example, the genotype of a particular cell or the biological description of that particular coat pigmentation as a common phenotype. As we have explained before, *(i)* makes reference to tokens, or instances, while *(ii)* makes reference to types, implying that these two terms (genotype and phenotype) are commonly used indistinctly to denote both tokens and types. Therefore, specific terms such as genotoken and phenotoken that would allow a clear distinction between tokens and types have only been mentioned sporadically in discussions about the gene and genotype ontology, without being considered a main issue. These few sporadic mentions are of note for two reasons. First, they illustrate very clearly that until now, it has not been possible to discuss type-token relations without actually talking about genes rather than talking about genotypes. In fact, in most cases, a clear distinction between

genotypes and genotokens versus gene*t*ypes and gene*t*okens is missing, and the same author (Lewontin) may use genotokens to denote both genes and genotypes. This is the case of this example by Sober and Lewontin (1982, p.165, note 4):

> Models of selection do not concern single organisms or the individual physical copies of genes (i.e., genotokens) that they contain. Rather, such theories are about groups of organisms which have in common certain genotypes. [...] Selection theory is about genotypes not genotokens.

Second, these examples show the potential of the type-token framework to uncover less explored aspects of the genotype-phenotype theory, anticipating certain unique features of genotypes that we will develop in the subsequent sections. Indeed, most of these explicit mentions of the type-token relation are used to discuss agency in the context of inheritance, in attempts to define which are the entities that evolve and how the continuity of genetic identity is ensured. To better understand these discussions, we can group them under two general points of view:

(i) An informational perspective, started by Williams' gene-centric views (Williams 2008) and continued more recently by Haig, Doolittle, and Inkpen (Haig 2014; Doolittle and Inkpen 2018). Based on Dawkins' early works (Dawkins 1976), these authors developed the idea of an immortal pattern as the genotype and its ephemeral avatars as their corresponding genotokens. According to Griesemer, this idea of immortality has its roots in *molecular Weismannism*, which derives from E. B. Wilson's misrepresented version of Weismann's theory of the continuity of the germplasm (Griesemer 2005). Although Haig, Doolittle, and Inkpen actually talk about genes, for these authors the continuity in the genotypic identity of genotokens would reside in an abstract (and immortal) informational pattern, disregarding the material genotokens themselves. In Haig's words (2014, p. 679, bold added by us):

> Material genes are physical objects but informational genes are the abstract sequences of which material genes were temporary vehicles. Material genes were identified with gene tokens and informational genes with gene types, but this is not quite right if 'type' is interpreted as a material kind. Sense DNA, antisense DNA, RNA and protein all represent an informational gene but are not molecules of one kind. **Continuity resides in the recursive representation of immortal pattern by ephemeral avatars**.

(ii) A genealogical and materialistic perspective, exemplified by Griesemer (Griesemer 2005) and, to some extent, by Wimsatt's criticisms (Wimsatt 1980) to the works by Williams and Dawkins referenced above (Dawkins 1976; Williams 2008). Griesemer emphasizes the materialistic aspect of inheritance and evolution. He does so by remarking the semiconservation of genotokens in DNA replication (Meselson and Stahl 1958) where one strand is conserved from the mother cell and one new is synthesized, making a half-new half-old DNA copy, therefore demonstrating that there is a material overlap, a persistence, with the next generation. In Griesemer's own words (2005, p. 76):

> Genotokens and phenotokens are inherited, but they are not the same tokens as the ones the parent inherited from its parents. The role of these tokens has been severely obscured

> by the obsession of transmission and molecular geneticists with analogies to the concept of information. [...] Persistence into the next generation (molecular, organismal, or whatever) is all that is required to make material overlap a property of the evolutionary process. Evolution is a material process, not a relation among abstractions in which biologically structured matter is accidental.

Thus, there would be a conservation of the same genotype identity based on a genealogical continuity between successive genotokens. This does not imply that there cannot be errors during replication. The resulting two genotokens (daughter DNA) from one genotoken (mother DNA) will always be slightly different, as a completely perfect molecular copy from a genotoken is almost impossible and different processes, such as recombination and repair, will introduce modifications. This point is important to understand the identity of the type from different genotokens, as this identity will eventually be eroded during evolution. The existence of absolutely perfect and identical clones of genotokens (which are probably impossible in the case of very large genomes) would be a transient stage that could only last for a limited number of generations, as mutations will inevitably appear.

To sum up, these two different positions make explicit the reference of the identity of genotokens. In the *(i) informational identity*, the reference is the parity of information represented by the abstract immortal pattern. In the *(ii) genealogical identity*, the reference is the genealogical line of descent of the genotokens, e.g., between mother and daughter cell, being possible to trace back a material overlap through generations. It is important to emphasize that in the materialist view of the second perspective, genotokens are the agents of the reproduction of their own identity by virtue of the uninterrupted stream of genotoken reproduction. By contrast, from the first perspective it would follow that the subject or agent of the action of reproducing an identity is the information extracted from the correlations between two or more genotypes.

In our view, from Griesemer's perspective it could be possible to define a type-token relation for genotypes independently of that of genes and of the translation of these genes into particular phenotypic outcomes or traits. Thus, as we will see below, it can constitute a new base to define genotypes versus genetypes and phenotypes. Under our perspective, genotypes as classification of whole inheritable structures could demand the same right of being natural kinds as biological species. We understand natural kind as a way of classifying individuals by scientifically discovering the structure of the natural world. By contrast, as we will show, genetype and phenotype are created (and limited) by human cognition, perception, cognition, and interests. Current discussions about natural kinds are succinctly explained in a recent work by Bird and Tobin (2022).

By contrast, the alternative perspective depends on the transmission of information, not only through generations but also when it is translated to the corresponding phenotypic traits: two structures that are completely different from a material perspective, protein and nucleic acids, are considered to conserve the same immortal informational pattern. In other words, an informational view of genotypes cannot be conceived independently of genes. As we pointed out at the beginning of this section, gene and genotype concepts have been traditionally intertwined, without

making explicit what is the difference between a gene as a type (classification of genes, what would be called genetype, as we previously mentioned) and the sum of all inheritable factors of an organism (genotype). In the next section, we will explore the difficulties of trying to individuate (*tokenize*) genes, showing that the type-token relation of genotype cannot be simply derived from genes.

3 Genotype and Genotoken Distinction

Genotype was originally conceived as "the sum total of all genes in a zygote or a gamete" (Johannsen 1911), suggesting that the genotype, as a type concept, is a class of its definite set or collection of genes (see types as sets (T1) in Box 1). For Johannsen, gene was a mere conceptual agreement to denote a hereditary factor, an individual cause of an individual trait, and he explicitly refused a material representation of genes or genotypes. This abstract conception changed completely in the subsequent decades when genes were proposed to have a material localization, leading to the *tokenization* of genes (and therefore genotypes) in the molecular content of chromosomes. This resulted in a reductionist nucleic acid-centric perspective, in contrast to the more pluralistic views of some early geneticists such as Johannsen and even Morgan, who wanted to encompass all possible sources of inheritance (Morgan 1917). But despite these changes, the view of genotypes as the collection of all the hereditary factors (the genes) has remained as the framework for the type-token relation of genotype. Thus, the individual instances of genotypes, the genotokens, have always been defined through the genes or, more precisely, following our initial quote by Lewontin, on the *postulated state* of the genes (Lewontin 1992, p. 137).

However, we argue that genes cannot be the ontological support for defining the genotype concept. Genes form their own types, in the same way as trait-tokens constitute phenotypes. Thus, we propose that the classification of genes as types has to be differentiated from that of genotypes. This difference between genes as tokens and genes as types is very rarely made explicit, with few examples where the type classification of genes is distinctively dubbed genetype (Chakrabarty 2010). Nevertheless, a type-token distinction in genes is implicitly acknowledged when genes are referred by their gene names and gene symbols out of their particular material context, for example, when genes are classified in bioinformatic databases and repositories. These bioinformatic genes are classified as types, being their referred tokens (genetokens) molecular segments of DNA localized in certain regions of chromosomes. And there we have to confront a fundamental problem: the challenge of individuating genes, that is, to precisely define the limits and locations of each of the genes contained in the genome.

3.1 The Classification of Genes Cannot Be the Framework for Genotoken Classification

To deal with the individuation of genes and clarify the distinction between genetype and genotype, we first have to take into account that there is more than one sense of gene coexisting in the scientific literature. Thus, we will review the three main and most widely used meanings of the term gene, as defined by Griffiths and Stotz (2013), to assess if they are suitable to define genotype as the type of genes, that is, if they are really able to unambiguously individuate genes. Furthermore, we will also discuss if these definitions of genes are insufficient to cope with all the potential hereditary factors of an organism. These factors can include multiple structures that are not nucleic acids but are nevertheless inherited and responsible for observable phenotypic traits. These can include cell membranes, which Cavalier-Smith termed genetic membranes because they self-template their own structure in each cell cycle, where they reproduce relative spatial dispositions of their building blocks, the phospholipids, in each new successive membrane copies (Cavalier-Smith 2004), as well as other self-guiding reproducing structures such as centrosomes or certain histone modifications (Ahmad and Henikoff 2018; Petryk et al. 2018; Yu et al. 2018; Nabais et al. 2020).

Definition 1 Functional Gene This is actually the original definition of genes, as they were first conceived as causes of previously and well-defined phenotypic traits, that is, as genetically segregating parts of the genotype of an organism (Johannsen 1911). Gene was considered equivalent to the *norm of reactions* (*Reaktionsnorm*) of Richard Woltereck (1909), namely, those hereditary factors susceptible to be selected from environmental conditions. In this sense, the norms of reactions were the characteristic response pattern to variation in environmental conditions (Falk 2009). This sense of gene has the advantage of bypassing the problem of relating genes to any material instance in particular; thus, there is no need to directly observe or identify genes within an organism. Therefore, using genes as reaction norms is still quite useful in all those situations where the focus is more on the phenotypic outcomes and adaptive value of the genes than on the material genes themselves. Also, in cases where genes have not been characterized at the genomic or molecular level, their existence can still be inferred by classical Mendelian approaches, that is, by observing the same potential phenotypic developments in individuals that are genealogically related and genetically identical. This entails that the individuality of a gene as an instance is created from and is subsidiary to a previous definition of a trait. A trait such as large seed size or alcohol production is related to one *gene*, creating a virtual 1-to-1 relation between a classified biological function (phenotype as large size) and its inherited reaction norm, termed *gene* (Griffiths and Stotz 2013). Thus, only after a phenotoken is well known and categorized as a phenotype that it makes sense to conceive a hereditary factor, termed gene (functional gene), whose actual material reference would only have to fulfill two requisites: (a) the gene has to be inside the phenotypic body of the biological individual being studied, and (b) it

has to be inherited through generations. Thus, this definition has the advantage that it could (in principle) be applied to other systems of inheritance that are not based on nucleic acids, consistent with the non-reductionist views of Johannsen and Morgan. Nevertheless, the type-token relation does not work properly with this sense of gene because there are no tokens of genes to take into account. This definition does not individuate a material reference of genes, only functional relations between undefined *hereditary factors*. Functional genes are present somehow within the organisms, but their precise material localization is not of particular relevance and is not conclusively specified. For instance, in Mendelian genetic maps the locations of genes do not correspond with actual physical distances in the DNA, but to nonmaterial relative distances with respect to other functional genes inferred from recombination frequencies (which are deduced from phenotypic observations). Thus, the identity of functional genes derives from the previously known traits. This way, phenotokens are the only defined tokens for functional genes.

Definition 2 Molecular Gene The molecular gene refers to the physical entity constituting a genotype part, e.g., a segment of a genome sequence that encodes a phenotypic trait or function, so genes would be like a string of beads in a linear molecule. The identity of such a gene is mixed because: *(i)* it still depends on a phenotype to be defined, being thus the gene equivalent to Mendel's alleles, and *(ii)* it depends on its specific sequence disposition on a genome sequence. In the initial years after the discovery of the genetic code, the definition of molecular gene seemed relatively straightforward, as the sequence of nucleotides encoding the synthesis of a protein product (i.e., a phenotypic trait but at the *molecular level*), with a 1-to-1 relationship between genes and proteins. However, this became increasingly problematic: first, with the discovery of introns and alternative splicing, implying that protein-coding genes were split into several fragments that could be combined in multiple manners (Berget et al. 1977; Chow et al. 1977a, b; Berk and Sharp 1977; Gilbert 1978), and second, with the growing catalog of noncoding genomic regions (such as the wide diversity of noncoding RNAs and cis-regulatory elements) that were nevertheless responsible for the expression and development of phenotypic traits (i.e., *noncoding* genes) (Cech and Steitz 2014). Finally, the characterization of an increasing number of genomes from a wide diversity of organisms, has revealed that all these elements, coding and noncoding, are intermingled in an extremely complex and interconnected manner within vast genomic regions that do not seem to have any obvious phenotypic impact (the so-called junk DNA), where coding exons constitute a tiny fraction (about 1% in the case of the human genome (Venter et al. 2001; Lander et al. 2001)). Furthermore, gene expression emerges from these DNA elements in a nonlinear three-dimensional (3D) manner: many noncoding regulatory elements cannot be ascribed to a single gene, since they regulate the expression of multiple genes, genes that can be located at very long distances from the perspective of a linear molecule or even within different genes and different chromosomes, making it impossible to delimit unique linear segments to define genes (Lettice et al. 2002; Spitz et al. 2003; Calhoun and Levine 2003; Markenscoff-Papadimitriou et al. 2014). In conclusion, the individuality of molecular genes cannot be well

established, and there is currently no clear consensus about how to delimit genes molecularly in a nonarbitrary and unambiguous manner. Finally, the definition of molecular gene was devised exclusively for nucleic acids (and, as explained before, from a particularly protein-centric and non-3D view of nucleic acids), and in its current form it is difficult to imagine how it could be applied to other hereditary structures such as cell membranes.

Definition 3 Informational Gene As genotype has been identified with genome sequence, it has been easy to match the ontology of the molecular gene with a third new sense: the informational gene. The informational gene refers to the information that can be extracted from a particular segment of the molecular genome sequence. Thus, in contrast to the molecular one, the informational gene can be reproduced in media that are different from its native nucleic acid molecules: the same genetic information would also be present in the amino acid sequence of the corresponding protein or even in artificial media such as a computer server. DNA and RNA nucleic acids are always composed by a group of just four different nucleobases, adenine (A), guanine (G), cytosine (C), and thymine (T, which in the case of RNA is replaced by uracil (U)). In this sense of gene, the specificity of the ordered sequence is presented as a cause of phenotypic variations. This ordered sequence of the same four nucleobases presented in any living being can be compared perfectly with the use of digits and letters in computation and written communication. In this sense, the genetoken is the particular and discrete fragment of these sequences such as ATGGTCTAA, and the genetype is the class of this sequence, *in the same manner* of the type-token relation in linguistics with our example of the type-word *rose*. This comparison between the information of the genome and the information related with human communication relies on encoding systems, such as the genetic code, that literally translate nucleic acids to proteins. Therefore, it works relatively well in the case of protein-coding genes, but defining information from other noncoding genomic regions in the same DNA is far more challenging and seemingly impossible to analogize with human communication systems. Additionally, the ordered sequence of nucleobases is just one of the inheritance systems of a living being, and similar to what happened with definition 2, it is not clear how sequential information could be applied to other hereditary structures such as genetic membranes, i.e., those membranes that are inherited independently of chromosomes. The complexity of the composition of cell membranes is still barely understood, mostly because the observation of the spatial arrangement and disposition of phospholipids is far more changeling than the much more stable and linear structure of genome sequences, making it difficult to conceive it from an informational perspective. In sum, it is not possible to reduce the components of the class of phenotypes to merely the ordered sequence of nucleobases (genetic information).

These three senses of gene concept live together in scientific literature and practice because they do not contradict each other: they make reference to different perspectives in biology (populations, molecular biology, bioinformatics). The main problem is that these definitions do not necessarily agree on their individuated genetoken, that is, the genes delimited by these definitions using genetic mapping,

molecular techniques, or bioinformatic approaches may not match (see for instance the examples in Table 1). Furthermore, all attempts to individuate genes have always been dependent on a previous individuation of a phenotoken, like readily observable morphological traits in the case of early geneticists, detection of molecular phenotokens of proteins and RNA transcripts, or bioinformatic evidence of the presence of a sequence with the potential of being translated into a protein. Thus, because of these problems, genes cannot be unambiguously defined as the individuated tokens of genotypes. In the next section, we will explain the core of our proposal to overcome these problems by viewing the material reference of a genotoken as a full and integrative structure inherited in each cell reproductive cycle. In this manner this material reference can be clearly distinguished as a unit regardless of the biological perspective, as in the case of a whole genome sequence or a whole genetic membrane, where their individuations are independent from external observations as they are granted by the cell's own reproduction.

3.2 Rethinking Genotypes as Self-Templated Replication Processes

We propose here a fourth reference for the genotokens of the genotype concept that complements the previous three: self-templated reproduction processes. In contrast to the previous ones, this new perspective focuses exclusively on the classification of entire genotokens as units, and it is the only one that does not depend on a previous definition of gene. Therefore, it is also independent from a corresponding individuation of phenotypes. Of course, our new proposal to define genotypes by focusing on whole genotokens rather than on ambiguously individuated parts (genes) could also be used to establish artificial kinds from a molecular or informational perspective, in a similar manner as the classification of molecular and informational genetypes. Given that genotypes and their identity are connected by uninterrupted genealogical processes, there are situations where, for pragmatic reasons, it is useful to define discrete genotypic identities by establishing artificial classifications and limits. Nevertheless, we think that the great potential of our proposal is that it shows that genotypes can be considered as natural kinds. This new reference is constructed upon four characteristics of genotokens that are implied in their classification as genotypes: *(i)* self-templated replication, *(ii)* maintenance of the same material nature, *(iii)* material overlap of genotokens, and *(iv)* integrative conservation of a whole genotoken as a unit.

By **(i) self-templated replication**, we mean that the copy of a genotoken's own structure is determined by the same genotoken during its own reproduction process. That is, the genotoken serves itself as a guide or template to generate a new genotoken, and this self-copying process is performed by the very same living system in which the genotokens are integrated. This does not mean that the genotoken has to carry out and catalyze the replication/reaction process alone.

Table 1 Genetype, genotype, and phenotype frameworks and their corresponding tokens

Type		Token definition	Examples of tokens	Kind relation of tokens (T2)
Genetype[a]	**Reference 1: Functional gene**	An inferred genetic cause of a phenotypic trait, mapped to a chromosomal location only in relative genetic terms (regardless of its precise molecular limits, that are either unknown or not pertinent)	Gene (as a token): *Iroquois* (*Iro*)[b] locus of a *Drosophila melanogaster* fruit fly	**Artificial kind**: Genes are classified as types by humans by using the characteristics of its associated phenotypes
	Reference 2: Molecular gene	A nucleotide segment within a DNA genomic molecule that gives rise to the expression of a molecular trait (RNA, protein)	Gene (as a token): Genomic region of 190 kb in chromosome 3 L of a *D. melanogaster* fruit fly, containing the transcriptional unit and cis-regulatory elements of *araucan* (*ara*)[b], one of the three *Iro* gene duplicates	**Artificial kind:** Genes are classified by humans by using the characteristics of the molecular structure that is transcribed to RNAs (and proteins, when coding)
	Reference 3: Informational gene	An ordered sequential pattern that can be abstracted as information that is transmitted and translated into a trait	Gene (as a token): Annotated nucleotide sequence information encoding the homeobox transcription factor protein ARA[b] of a fruit fly, represented in the form of text, hardware (e.g., silicon), software (e.g., binary), etc.	**Artificial kind**: Genes are classified by humans by addressing their particular sequences
Genotype	**Reference 4: Self-reproduction processes**	Whole inherited structure that self-reproduce every cellular generation	Genotoken: DNA molecules that constitute the whole genome of a yeast cell Genotoken: Whole cell membrane structures (*membranome*[c]) of a yeast cell	**Natural kind**: Genotokens are classified as genotokens independently of any interest or action of human beings
Phenotype	**Context-dependent references** (molecular, physiological, morphological)	Perceivable trait or biological function at all possible levels (molecular, metabolic, morphological,	Phenotoken: Plumage coloration of the wing of a bird Phenotoken: Hexose transporter activity of GHT1 proteins in a	**Artificial kind**: Phenotokens are classified by humans through the assessment of their similarity of

(continued)

Table 1 (continued)

Type		Token definition	Examples of tokens	Kind relation of tokens (T2)
		physiological, behavioral, etc.)	*Schizosaccharomyces pombe* yeast cell	observable characteristics by an agent that is external to the phenotokens (i.e., humans)

[a]We use here genetype instead of genotype because current frameworks to define genotypes are subsidiary to the different definitions of genes and therefore cannot be directly compared to our genotype-genotoken framework since they do not classify integral inherited structures

[b]We have used the examples of the genes belonging to the Iroquois gene complex, the tandemly duplicated genes *araucan* (*ara*), *caupolican* (*cau*), and *mirror* (*mir*), to show how the three definitions of genes can lead to different gene individuations. The functional gene identified from a mutant phenotype (*Iro*, causing the loss of bristles rows in the flies' notum) was actually caused by a loss of function of shared long-range regulatory elements affecting several protein-coding genes, *ara* and *cau*, whose separate mutations do not cause visible phenotypic effects (Cavodeassi et al. 2001). Because of this complex arrangement of shared cis-regulatory elements, to individuate the molecular segment containing all the regulatory elements of just one of these genes such as *ara*, we have to consider a 190 kb genomic region that includes also *cau* and *mirr*, together with an unrelated protein-coding gene, *sowah*, and several noncoding RNA genes (Maeso et al. 2012)

[c]By membranome we mean here the whole configuration of the genetic membranes as it is inherited in reproductive cycles. This has been first coined by Cavalier-Smith (2004)

This may have been the case in the early stages of the origin of life with RNA molecules with ribozyme activity. In that case, self-templated replication of the genotoken would refer not only to this capacity to serve as its own guide or template but also to the capacity of the genotoken's own structure to act as the replicator and catalyst of a self-replication reaction. However, in current living systems that have DNA genomes as genotokens, the replication of a new genotoken-DNA-molecule depends on the presence of a template-DNA-molecule but other structures, such as DNA polymerases and all the other proteins that are part of the replication complexes, are the ones that carry out the synthesis of the new genotoken. Nevertheless, these phenotokens that take part in the self-templated replication process of the genotokens are always also part of the biological identity of the living being to which the genotoken belongs, either as direct phenotypic expressions of the genotoken that is being replicated or of other genotokens within the same organism (Table 1). In our opinion, cases like viruses, parasites, and other complex associations between several organisms such as holobionts, where the self-templated replication of genotokens from viruses and other microbial entities may be carried out using phenotokens of the *host*, do not constitute an exception to our theory. In these cases, we consider that the phenotokens and genotokens involved are in fact forming a complex biological individual with more than one genotypic identity (e.g., holobionts). Finally, comparing DNA self-templated replication with that of genetic membranes is unambiguous: newly synthesized phospholipids and proteins are ordered by contact with the ones already assembled in the membrane that will be

eventually divided and duplicated, as in the case of DNA replication. Similarly, this does not happen in an autocatalytic manner, and genetic membranes also make use of diverse protein machinery to carry out their replication processes (Cavalier-Smith 2004).

By **(ii) same material nature**, we refer to the fact that genotokens conserve the same kind of precursor materials in their reproduction. These material components are recruited during self-templated replication to recreate in the new genotoken copies the specific three-dimensional configuration of the preexisting genotokens. Taking DNA replication as an example, it can be observed that every replicated copy is made by using the same type of molecules as building blocks (nucleotides) that the copy that is being replicated did. The building blocks of genotokens are also natural kinds because they belong to the same chemical compounds: nucleotides in DNA or phospholipids in the case of genetic membranes (for a discussion of chemical compounds as natural kinds, see Bird and Tobin (2022)). This is in contrast to the transmission of information from DNA, when one ordered structure is translated to another one of different kind, i.e., a phenotoken, such as a protein, that results from DNA and RNA expression. Here, the concept of information is relevant to show the analogy between two material structures of different type connected by relation of material translation from genotoken to phenotoken: nucleic acid patterns are translated by ribosomes to form amino acid patterns using the equivalence of the genetic code, always following the pattern of three nucleotides (codons or triplets) for a single amino acid. This is a drastic and irreversible change in the material nature that implies a different type (genotype to phenotype): as soon as the patterns change from nucleic acids to amino acids, there are immediate reactions where the newly formed structure will be a protein folded by multiple environmental factors. On the other hand, replication and reparation processes of genotokens always conserve the nature of the material structures, being these processes aimed at copying the previous model and to maintain its material integrity, respectively. Importantly, the maintenance of the same material nature is also a property of additional inherited structures such as genetic membranes, histone modifications, or centrosomes. In genetic membranes phospholipids are also recruited to the preexisting membrane, so its specific configuration and polarity can be maintained in every cell division (Carlton et al. 2020). Unfortunately, in contrast to DNA, the molecular processes responsible for copying the spatial configurations of the components of membranes and other structures, such as centrosomes and modified histones, are still poorly understood. So far, the molecular assembly of these hereditary systems has not been framed yet as possible processes of self-templated replication of genotokens. We believe that investigating these assembly processes under this new perspective will be key to their understanding.

The **(iii) material overlap** of genotokens can be seen as a consequence of *(ii)*. This is the core of Griesemer's critique of the abstract interpretation of genotypes (Griesemer 2005) and is clearly illustrated in the case of the *semiconservation* of DNA strands during replication (Meselson and Stahl 1958), which was a key aspect in the elucidation of DNA as a hereditary structure. This means that the double strand of DNA is separated in two to recreate the same double helix by the overlap of new

nucleobases with the old ones. Thus, the path of evolution is not a path of genotypes linked together by an identity that is communicated between generations as information, as it is implied in the metaphor of genes using material genomes simply as *vehicles* (genotype as information). Rather, this processual identity is based on the fact that there is a material connection between the replicating genotoken that acts as a template and the resulting genotoken copy. We propose here that this is a fundamental property of hereditary structures: they reproduce their structure by recruiting new building blocks and overlapping them with the old ones to replicate their configuration. In fact, in addition to DNA, all other inherited structures known so far fulfill this criterion, as in the case of genetic membranes, where new phospholipids contact the old ones expanding and growing the ordered structure of the membrane which will be divided into two membranes during cell division (Cavalier-Smith 2004), centrosomes, where a daughter centriole buds from the mother one (Nabais et al. 2020), or modified histones (Ahmad and Henikoff 2018; Petryk et al. 2018; Yu et al. 2018).

Finally, with **(iv) the integrative conservation of the whole genotoken as a unit**, we want to emphasize the unity of genotokens versus the unity of genes. Genes within a genome sequence are not replicated and inherited per se as molecular units, but as a consequence of being indissoluble fragments of the much larger and highly complex DNA molecules that together with many other structures form the chromosomes. Genes are just non-physically separated parts of the whole genome sequence, and from the perspective of the replication machinery and the subsequent inheritance during cell division, the individuality of these molecular fragments (genes) makes no difference. Therefore, the unit of the whole genome genotoken is recreated in every cell generation regardless of any gene under any of its definitions. We think this is also the case for other hereditary structures, such as genetic membranes or centrosomes. One important consequence of the integrative conservation of the genotoken as a unit is that the type relationship of genotype between its genotokens does not depend on an external categorizing agency. This is so because the conservation and integrity of the type are self-sustained by the same entity being duplicated: a cycling cell and its components.

By accepting this definition for genotokens being categorized as genotypes, two immediate benefits follow:

1. Our framework opens the possibility of categorizing other hereditary structures as genotypes in addition to DNA and RNA genomes. In this way, the different structures that can be inherited between generations such as genetic membranes, modified histones, centrosomes, and prions could potentially be referred to as genotokens by looking for structural identity relationships that depend on their copying processes rather than on the possibility of being translated as information (a translation that so far is exclusively restricted to protein-coding genes and therefore fails to account for additional genotype-to-phenotype relationships). Thus, here we propose that this identity would be grounded on the self-templated replication of the genotoken structure to be reproduced in the successive generations.

2. It offers an alternative to the informational gene perspective (Definition 3) to address the identity of genotypes. By basing it on the material genealogical link that connects successive genotokens, the identity is not the abstract correspondence between two genotokens and their computational representation obtained by sequencing them. Instead, the identity is based on the process of reproduction by materially overlapping molecular building blocks to form genotokens. Mutations and other sources of alteration of the copied pattern happen, being regularly repaired by specific molecular processes. But these repairing processes do not rely on an informational framework, but either on overlapping processes equivalent to all genotoken reproduction processes or on random stitching of broken parts (which in the case of DNA would correspond to homologous versus nonhomologous repairing mechanisms, respectively). That the molecular ordered sequence can be abstracted away from its material structure and represented by external media, such as text or binary code, does not imply that genotokens themselves are self-identified by fidelity to anything outside their own material presence, while this kind of *external* fidelity (i.e., independently of the media) occurs in every copying process of information as information.

In conclusion, this new way of using the genotype concept, which in our view should encompass all the potential hereditary factors of an organism, allows the inclusion of all the inherited structures that a cell copies during its reproductive cycle. Under our perspective, as long as they reproduce by self-templating, hereditary systems that have been until now termed as nongenetic or as extended or epigenetic inheritance can be considered an integral part of the genotypic identity of an organism. Although further investigations will be required to better understand their reproduction processes, our extended view of the genotype could include some of the structures mentioned earlier (genetic cell membranes, centrosomes, modified histones) as well as other higher-order systems, such as learned behaviors within animal social groups and certain supra-organismal associations. Some of these structures may be more amenable to be represented as sequential information, such as the stable and ordered succession of the nucleobases composing genome sequences, while other structural patterns are more difficult to address as information, such as the dynamic structures formed by the union of phospholipids and proteins in genetic membranes.

4 Phenotype and Phenotoken Distinction

Once we have shown our perspective on the type-token relation of the genotype concept, we come back to the genotype-phenotype difference and focus on the classification of phenotokens in phenotypes. Following the aforementioned definition by Lewontin, phenotypes are the "observable physical qualities of the organism, including morphology, physiology, and behavior at all levels of description" (Lewontin 1992). In a type-token distinction, this implies the phenotype is a relation

of *kind* (see T2 in Box 1), being the phenotype a group of qualities or states and not a group of tokens (as in the case of T1, Box 1). As far as we know, phenotype is never used as the group of phenotokens of organisms, but as a model of physical qualities and phenomena.

The qualities that phenotokens represent are the result of a particular combination of genotypic expression and environmental conditions. These particular combinations are repeated in every life cycle, generating relations of identity between phenotokens. What is understood as *one* phenotoken is flexible enough to adapt to any biological feature: the chemical structure of a protein synthesized by a ribosome following the pattern of a mature mRNA but also the microenvironment created by a cell membrane, the metabolic rate of a mouse, or the eye pigmentation of a fly. These are examples of phenotokens, although materially speaking they are very different kinds of tokens. Furthermore, the growing advances in the technology applied to observe biological features are exponentially increasing the quantity of phenotokens to be classified (Nachtomy et al. 2007, 2009).

Compared with genotokens, which are self-defined individuals by means of their own self-templated reproduction, the individuality of phenotokens cannot be directly attributed to their own activity. The being of a phenotoken begins with genotypic expression. Genotypic expression is understood as the translation of patterns contained in genotokens into the configuration of a new product called phenotoken. The paradigmatic genotypic expression is the translation from a nucleotide sequence (a genetoken) to an amino acid chain (a phenotoken). Nevertheless, one of the main consequences of expanding the genotype category beyond protein-coding genes and, in our perspective, even to other inherited structures like genetic membranes is that it consequently extends what is understood as a phenotoken. All inherited structures known so far have distinctive functions that characterize them. However, these functions have never been interpreted as the outcome of the genotypic expression of these non-DNA structures. Although the description of phenotypes expressed from other genotokens different from genome sequence is far beyond the scope of this chapter, as a starting point we can mention as an example the review made by Moog and Maier, where they explain how membranes configure the intramembranous spaces of the cell, separating plasmatic from non-plasmatic phases and creating microenvironments that are crucial to the development of biological functions (Moog and Maier 2017). Furthermore, more recently Nwamba has reviewed the functionality of genetic membranes to be understood as genetic expression (Nwamba 2020).

Although any phenotoken is a natural product of genotypic expression (in a particular environment), their individuation as a token and their identification in a type-category is artificial and intentional. This means that there is a selection of the limits of what we, as biologists, want to conceive as a unit, as well as a selection of the qualities of interest, which are both the base of the classification of the tokens in types. This is why we propose that phenotypes are formed as an artificial kind relation between phenotokens. Genotypes can be considered natural kinds if they are defined by the classification of genotokens that are replicated by self-templating and, therefore, categorized in the act of their reproduction. By contrast, phenotokens

as products of genotypic expression are classified by an intentional classification linked to perceptive and cognitive conditions of us, humans, and to the pragmatic requirements of the process of classification of the different biological disciplines (see Table 1). For instance, the qualities of a phenotoken selected by a molecular biologist and how these qualities are classified into types follow the scientific framework of biochemistry, while the qualities of a phenotoken selected by a zoologist would generally be more related to macro-physical qualities and based on morphology and physiology. Nevertheless, the fact that we are considering phenotypes as artificial kinds does not mean that phenotypes are spurious categories in respect to genotypes. Indeed, phenotypes are an essential theoretical tool for the description, quantification, and classification of all biological entities.

5 Concluding Remarks

What exactly are the genotokens of the genotype of a living being is, under our view, a crucial question that has so far been overlooked. The main consequence of not properly defining the tokens of genotypes is that the definition of genotype as a concept has remained elusive and has always been subsidiary to that of genes. In this work, we have proposed a new perspective on the genotype concept that differentiates it from the phenotype concept. We have addressed this question through two complementary aspects.

First, by pointing out that the genotype concept should classify whole inherited structures, not genes, as genes, under any of their currently accepted definitions, can only refer to partial and arbitrary segments within these hereditary structures. Thus, genes are weakly defined both as individuals and as tokens. In contrast, by considering integral inherited structures as tokens, our proposal can encompass all the potentialities of the genotype. As an example, in the case of genomic DNA, this allows the consideration of all of its molecular characteristics, many of which, such as 3D conformations, are not easily included or represented by genes or by informational representations as mere sequences. Furthermore, this opens the possibility of recognizing other potential genotokens such as genetic membranes and centrosomes that could not have been identified as such under a gene-centric paradigm, as they cannot be easily conceived as composed of multiple *genetic loci*.

Second, we have defined the properties of these integral inherited structures based on how these tokens autonomously recreate their own individuality and identity. With this, we have proposed a novel framework that consists of four characteristics: self-guided replication of integrative and materially overlapping genotokens. These four characteristics offer an alternative to the interpretation of the genotype as sequential information, as they can be potentially fulfilled by other biological structures beyond DNA and RNA genomes.

In contrast, in the case of the phenotype concept, our discussion has been limited to point out that what is conceived as a phenotoken and its classification into phenotypes is context dependent, i.e., adapted to the praxis of each biological

discipline. Therefore, we consider that this is an eminently pragmatic concept that can maintain its current meaning and uses.

Our new perspective suggests that the inherited structures of an organism, such as the so-called epigenetic inheritance, genetic membranes, or even more macroscale structures like holobionts, can constitute the genotypic identity of an organism. Therefore, in our view, the fact that these hereditary systems are not based on nucleic acids does not imply that they are nongenetic. This only means that they are non-genomic, and we think that the use of the expression *nongenetic* to refer to hereditary structures should be avoided.

In conclusion, we hope that our work can pave the way to a more complete and less reductionist view of heredity, allowing the additional inherited structures to be fully integrated within the genotype-phenotype framework.

Box 1 Characterization of Type-Token Relations

The type-token difference is a perspective that comes from linguistics and philosophy of language. When it was defined for the first time, the type-token distinction was related to semantics (Peirce 1931-58), where different single-tokens such as words or ideograms were related by their meaning in types. Nevertheless, type-token differences can be applied to virtually any relations of identity between individuals, for example, biological individuals in species (Wetzel 2006). In our opinion, the main value of this distinction is that it establishes a relation of unity between the tokens and the type, making explicit the existence of the type as a unit of reference. The ways tokens are related define therefore the type concept, making potentially unlimited type variations. Lynda Wetzel offered a classification of type concepts in three general relations between tokens, which follows a different character to the individuality of the type. In brief, these are Type-Token 1 (T1), sets as wholes with a finite number of tokens; T2, kinds with explicit relations of group of characteristics between tokens; and T3, laws with made-up networks organizing tokens (Wetzel 2006). Though these three do not constitute an exhaustive list of all possible ontologies for type-token relations, they offer a good background to review their implications for the genotype and phenotype concepts.

T1: Types can be a *set* or collection of tokens. The relation between the tokens constituting a set can be natural or artificial, meaning that classification in sets can be based on a description of observations (e.g., natural kinds), or can be a made-up classification, like a chosen collection of numbers, or even a collection of different kinds of tokens, like a list of goods owned by a person (encompassing food, bikes, books, etc.). Types as sets have a definite number of tokens, being the sum of all tokens the components or parts of the type as a whole. The first definition of genotype given by Johannsen offers a clear example: the genotype as the sum of all the genes of an organism. This sum

(continued)

Box 1 (continued)
is the set of genes reproduced in each generation. Sometimes types as sets are viewed more explicitly as emerging individuals from its own tokens, as in the classic example of David Hull who proposed the individuality of a biological type species as the entity delimited by the reproduction of their own organisms (Hull 1976, 1978). When type species are defined as an individual, it entails a delimited collection of tokens in discrete time and space coordinates. Our view of the genotype concept could follow an analogous perspective: the sum of all genotokens of a cell is a delimited collection of inheritable structures, which would include DNA, genetic membranes, and other epigenetic structures.

T2: Types can be *kinds*. This type-token relation is constructed on kind relations that do not necessarily imply a collection, a set, or a whole-part relation. Types as kinds are constituted as the group of characteristics that all the tokens belonging to that type must fulfill. These characteristics forming a type imply a description of a similarity between their tokens. As in the case of types as sets, the relations between tokens can be made up or based on natural kinds. Following Quine, it is important to note "kinds can be seen as sets, determined by their members. It is just that not all sets are kinds" (Quine 1969, p. 118). Conversely, not all kinds are sets, as there can be kinds that contain a potentially unlimited number of tokens, like in the kind *stars*, since it cannot be determined if there is a definite number of them in the universe. Types as kinds are commonly represented by an idealistic or paradigmatic token as an example that includes all the features described by the type, as in Plato's ideas where one idea is the perfect example of the imperfect copies. In our view, the genotype of a cell could be both considered as T2, as it has a relation of natural kind, and T1, as they also form a set of genealogically connected tokens.

T3: Types can be *laws*. In this case, types refer to the specific relations between tokens made by agreement/convention without implying the existence of any emerging individual as the whole from these parts; therefore, it is not possible to select any perfect example or holotype. Types as laws are precisely the way in which the type-token relation was conceived by Charles Sanders Peirce, the author who defined type and token for the first time (Peirce 1931-58). Peirce used type classification of signs as tokens to point out the semantic relations that classify these into types. A traffic signal, a Chinese ideogram, or a syllabic word as tokens constitutes the reference of semantic networks that classify the relation between tokens. These can be shared between different copies of tokens, but for type-token relations as laws, the quantity or the perfection of these copies is not relevant. Sometimes genotype and phenotype difference has been considered to be equivalent to semantic and informational channels, where genes are the possible meanings and the phenotypes are the expression through environmental channels of potential distortion (Smith 2000; Barbieri 2015). This theory is evaluated in our definition of gene (see Sect. 3.1, Definition 3).

Acknowledgments and Financial Support We would like to thank two anonymous reviewers for their constructive criticisms on the manuscript. We also thank Isabel Almudi for her help and comments. This work has been funded by the Spanish Ministry of Science and Innovation and the European Union (grants RYC-2016-20089, PGC2018-099392-A-I00, and PID2021-128728NB-I00 to IM) and by the *Fondo Europeo de Desarrollo Regional* (FEDER) and the *Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía*, within the operative program FEDER Andalucía 2014–2020 (*01-Refuerzo de la investigación, el desarrollo tecnológico y la innovación*, grant P20_00419 to IM).

References

Ahmad K, Henikoff S (2018) No strand left behind. Science 361:1311–1312. https://doi.org/10.1126/science.aav0871

Alberch P (1991) From genes to phenotype: dynamical systems and evolvability. Genetica 84:5–11. https://doi.org/10.1007/BF00123979

Barbieri M (2015) Code biology: a new science of life. Springer

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A 74:3171–3175. https://doi.org/10.1073/pnas.74.8.3171

Berk AJ, Sharp PA (1977) Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. Cell 12:721–732. https://doi.org/10.1016/0092-8674(77)90272-0

Bird A, Tobin E (2022) Natural kinds. In: Zalta EN (ed) The Stanford Encyclopedia of philosophy. Spring. 2022 Edition. https://plato.stanford.edu/archives/spr2022/entries/natural-kinds

Bonduriansky R, Day T (2020) Extended heredity: a new understanding of inheritance and evolution. Princeton University Press, Princeton

Calhoun VC, Levine M (2003) Coordinate regulation of an extended chromosome domain. Cell 113:278–280. https://doi.org/10.1016/s0092-8674(03)00309-x

Carlton JG, Jones H, Eggert US (2020) Membrane and organelle dynamics during cell division. Nat Rev Mol Cell Biol 21:151–166. https://doi.org/10.1038/s41580-019-0208-1

Cavalier-Smith T (2004) The Membranome and membrane heredity in development and evolution. In: Hirt RP, Horner DS (eds) Organelles, genomes and eukaryote phylogeny: an evolutionary synthesis in the age of genomics. CRC Press, pp 382–403

Cavodeassi F, Modolell J, Gómez-Skarmeta JL (2001) The Iroquois family of genes: from body building to neural patterning. Development 128:2847–2855

Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157:77–94. https://doi.org/10.1016/j.cell.2014.03.008

Chakrabarty P (2010) Genetypes: a concept to help integrate molecular phylogenetics and taxonomy. Zootaxa 2632:67. https://doi.org/10.11646/zootaxa.2632.1.4

Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977a) An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell 12:1–8. https://doi.org/10.1016/0092-8674(77)90180-5

Chow LT, Roberts JM, Lewis JB, Broker TR (1977b) A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. Cell 11:819–836. https://doi.org/10.1016/0092-8674(77)90294-x

Dawkins R (1976) The selfish gene. Oxford University Press, New York

Doolittle WF, Inkpen SA (2018) Processes and patterns of interaction as units of selection: an introduction to ITSNTS thinking. Proc Natl Acad Sci U S A 115:4006–4014. https://doi.org/10.1073/pnas.1722232115
Ellis BD (2001) Scientific essentialism. Cambridge University Press, Cambridge
Ereshefsky M (1998) Species pluralism and anti-realism. Philos Sci 65:103–120. https://doi.org/10.1086/392628
Falk R (2009) Genetic analysis: a history of genetic thinking. Cambridge University Press, Cambridge
Gilbert W (1978) Why genes in pieces? Nature 271:501. https://doi.org/10.1038/271501a0
Griesemer JR (2005) The informational gene and the substantial body: on the generalization of evolutionary theory by abstraction. In: Jones MR, Cartwright N (eds) Idealization XII: correcting the model: idealization and abstraction in the sciences. BRILL, pp 59–115
Griffiths P, Stotz K (2013) Genetics and philosophy: an introduction. Cambridge University Press, Cambridge
Haig D (2014) Fighting the good cause: meaning, purpose, difference, and choice. Biol Philos 29:675–697. https://doi.org/10.1007/s10539-014-9432-4
Hull DL (1976) Are species really individuals? Syst Zool 25:174. https://doi.org/10.2307/2412744
Hull DL (1978) A matter of individuality. Philos Sci 45:335–360. https://doi.org/10.1086/288811
Jablonka E, Lamb MJ (2005) Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life. The MIT Press, Cambridge (Mass)
Johannsen W (1911) The genotype conception of heredity. Am Nat 45:129–159. https://doi.org/10.1086/279202
Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062
Lettice LA, Horikoshi T, Heaney SJH et al (2002) Disruption of a long-range cis-acting regulator for shh causes preaxial polydactyly. Proc Natl Acad Sci U S A 99:7548–7553
Lewontin RC (1992) Genotype and phenotype. In: Keller EF, Lloyd EA (eds) Keywords in evolutionary biology. Harvard University Press, Cambridge, Mass
Maeso I, Irimia M, Tena JJ et al (2012) An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. Genome Res 22:642–655. https://doi.org/10.1101/gr.132233.111
Markenscoff-Papadimitriou E, Allen WE, Colquitt BM et al (2014) Enhancer interaction networks as a means for singular olfactory receptor expression. Cell 159:543–557. https://doi.org/10.1016/j.cell.2014.09.033
Meselson M, Stahl FW (1958) The replication of DNA in *Escherichia coli*. Proc Natl Acad Sci U S A 44:671–682. https://doi.org/10.1073/pnas.44.7.671
Moog D, Maier UG (2017) Cellular compartmentation follows rules: the Schnepf theorem, its consequences and exceptions: a biological membrane separates a plasmatic from a non-plasmatic phase. BioEssays 39:1700030. https://doi.org/10.1002/bies.201700030
Morgan TH (1917) The theory of the gene. Am Nat 51:513–544. https://doi.org/10.1086/279629
Nabais C, Peneda C, Bettencourt-Dias M (2020) Evolution of centriole assembly. Curr Biol 30:R494–R502. https://doi.org/10.1016/j.cub.2020.02.036
Nachtomy O, Ramati Y, Shavit A, Yakhini Z (2009) It takes two to tango: genotyping and phenotyping in genome-wide association studies. Biol Theory 4:294–301. https://doi.org/10.1162/biot.2009.4.3.294
Nachtomy O, Shavit A, Yakhini Z (2007) Gene expression and the concept of the phenotype. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci 38:238–254. https://doi.org/10.1016/j.shpsc.2006.12.014
Nwamba OC (2020) Membranes as the third genetic code. Mol Biol Rep 47:4093–4097. https://doi.org/10.1007/s11033-020-05437-z
Peirce CS (1931-58) Collected Papers of Charles Sanders Peirce, Volumes I and II: Principles of philosophy and elements of logic. 1985. Belknap Press of Harvard Univ. Press, Cambridge, Mass

Petryk N, Dalby M, Wenger A et al (2018) MCM2 promotes symmetric inheritance of modified histones during DNA replication. Science 361:1389–1392. https://doi.org/10.1126/science.aau0294

Quine WV (1969) Ontological relativity and other essays. Columbia Univ. Press, New York, NY

Sapp J (1983) The struggle for authority in the field of heredity, 1900?1932: new perspectives on the rise of genetics. J Hist Biol 16:311–342. https://doi.org/10.1007/BF00582405

Smith JM (2000) The concept of information in biology. Philos Sci 67:177–194. https://doi.org/10.1086/392768

Sober E, Lewontin RC (1982) Artifact, cause and genic selection. Philos Sci 49:157–180. https://doi.org/10.1086/289047

Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113:405–417

Taylor PJ (2018) An invitation to explore unexamined shifts and variety in the meanings of genotype and phenotype, and their distinction. Philos Theory Pract Biol 10. https://doi.org/10.3998/ptpbio.16039257.0010.006

Taylor PJ, Lewontin RC (2017) The genotype/phenotype distinction. In: Zalta EN (ed) The Stanford Encyclopedia of philosophy, summer 2021 edition. https://plato.stanford.edu/archives/sum2021/entries/genotype-phenotype

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351

Wetzel L (2006) Types and tokens. In: Zalta EN (ed) The Stanford Encyclopedia of philosophy, fall 2018 edition. https://plato.stanford.edu/archives/fall2018/entries/types-tokens/

Williams GC (2008) Adaptation and natural selection: a critique of some current evolutionary thought, with a new preface by the author. Princeton University Press, Princeton, NJ

Wimsatt WC (1980) The units of selection and the structure of the multi-level genome. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1980:122–183. https://doi.org/10.1086/psaprocbienmeetp.1980.2.192589

Woltereck R (1909) Weitere experimentelle Untersuchungen über Artänderung, speziell über das Wesen quantitativer Artunterschiede bei Daphniden. Verh Dtsch Zool Ges 19:110–173

Yu C, Gan H, Serra-Cardona A et al (2018) A mechanism for preventing asymmetric histone segregation onto replicating DNA strands. Science 361:1386–1389. https://doi.org/10.1126/science.aat8849

The Gene as a Natural Kind

Francesca Bellazzi

Abstract What is a gene? Does it represent a natural kind, or is it just a tool for genomics? A clear answer to these questions has been challenged by postgenomic discoveries. In response, I will argue that the gene can be deemed a natural kind as it satisfies some requirements for genuine kindhood. Specifically, natural kinds are projectible categories in our best scientific theories, and they represent nodes in the causal network of the world (as in Khalidi. Natural Categories and Human Kinds: Classification in the Natural and Social Sciences. Cambridge University Press, Cambridge, 2013; Khalidi. Synthese 195: 1379–1396, 2018; Khalidi. Are Sexes Natural Kinds, In: Dasgupta S, Weslake B (eds) Current Controversies in Philosophy of Science. Routledge, New York, 2020; Khalidi. Philos Sci 88:1–21, 2021). In Sect. 2, I will present a brief history of the gene and the controversy over its status. In Sect. 3, I will introduce the account of natural kinds considered in this paper. In Sect. 4, I will first present the relevant definition of genes and how they can be classified. Then, I will argue that the gene can be considered a natural kind as it satisfies the criteria for natural kindhood. Section 5 concludes.

Keywords Gene · Natural kinds · Projectibility · Causal nodes · Genetics

1 Introduction

What is a gene? Is it a natural kind, or is it just a tool for genomics? A clear answer to these questions has been challenged by the developments of genetics of (at least) the last 20 years (Griffiths and Stotz 2006, 2013). On the one hand, the identity relation between individual genes and precise stretches of DNA has proved impossible. This might support a deflationary or nominalist view of the gene (Griffiths and Stotz 2006; El-Hani 2007; Fogle 2010). On the other hand, some have been arguing that we should maintain a realist approach to the gene while understanding better the

F. Bellazzi (✉)
Department of Philosophy, University of Bristol, Bristol, UK
e-mail: francesca.bellazzi@bristol.ac.uk

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_12

cellular context in which it operates and embracing its complexity (El-Hani 2007, Griffiths and Stotz 2013, Bellazzi 2022).

The controversy is not easily settled, and one of the questions risen is whether the gene is a category "neat" enough to count as a natural kind or whether it is instead too ambiguous. Here, I will consider this question, and I will argue in favour of a positive answer as the gene satisfies some requirements for natural kindhood, following Khalidi's account of natural kinds (2013, 2018, 2020, 2021). These are genuine projectibility and being a node within causal networks. In a nutshell, the thesis defended here is that, under the considered approach to natural kinds, the category "molecular gene" used in scientific practice corresponds to a natural kind despite the complexity of the properties characterising it, and it captures some objective features of reality. This is a captivating case study because it illustrates how natural kinds can be found even within complex and highly interactional systems.

This inquiry has some conclusions of interest. First, understanding whether something is a natural kind or not is important because the naturalness of a given category can provide us with a further justification for why we can make more robust inferences from it. In doing so, the identification of something as a natural kind can support the justification of a theory that presents such a kind. Second, a natural kind is more than a theoretical entity whose properties are postulated for practical purposes, and this can direct research into discovering (rather than mere postulating) features about it. This supports the role that they have *also* in the process of discovering new information about such category. A natural kind is correspondent to something objective in the world, meaning that some properties could be discovered about it, and some could not. Lastly, and more generally, this project represents an instance in which biology and biological practice inform the philosophical debate on what counts as a natural kind. The structure of this paper will be the following. In Sect. 2, I will provide a brief overview of the history of the gene, in order to present the status of the controversy. Then, I will consider a definition of the gene within the contemporary debate. In Sect. 3, I will present the relevant account of natural kinds considered, as presented by Khalidi. In this framework, natural kinds are projectible categories and nodes in causal networks. Having provided the metaphysical ingredients, in Sect. 4 I will move on to consider whether the gene is a natural kind. I will argue that this is the case as it satisfies the criteria for natural kindhood aforementioned. Section 5 concludes.

2 A Brief History of the Gene

The last century has been rightly called *The Century of Gene* by Fox Keller (2000). After the disruptive discovery and spread of the theory of evolution, the twentieth century started with the aim of solving the puzzle of stability of traits and their transmission. In 1900, three journal articles were published by de Vries, Correns and von Tschermak exposing the laws of genetics, further developing and re-interpreting

Mendel's studies on hybridisation[1] (El-Hani 2015). It was clear that something was transmitted discretely from parents to offspring in a way that respected precise probability distributions. What was unclear was *the nature* of the entity under discussion. There were many candidates at the time, such as *gemmules* proposed by Darwin, de Vries' *pangenes* or Weismann's *determinants*. These were possible different entities that could have been transmitted respecting the laws of genetics (Fox Keller 2000; Beurton 2010; Falk 2010). In response to this conceptual unclarity, Johannesen decided to introduce the new term *gene* to disentangle the discussion and to facilitate scientific practice. In 1909, he presented the gene as what explains the transmission of traits from parents to offspring and should have been "free from any hypothesis", with no theoretical pre-assumptions. It was considered only a "convenient notational concept" (Falk 2010, 321). First introduced as an instrumental theoretical entity, one whose observation wasn't needed, but whose postulation could improve the explanations and predictions of a theory, the gene changed our way of doing biology.

Nevertheless, not all scientists were aligned with an instrumentalist view of genetics. With its development, genes started to have properties in their own right, even if they were detectable only indirectly. The tensions between concrete experimental demands and the specification of the phenomenon under consideration led towards a more specific material identification of the gene (Fox Keller 2000; El-Hani 2007; Falk 2010). In particular, Muller, a student of Morgan, supported a realist hypothesis of genes as material units, possibly chemical molecules, in disagreement with most scientists and his master (El-Hani 2007). This realist understanding supported the research for a material molecular basis of the gene. In the early 1940s, there was the rising of the one-gene-one-enzyme hypothesis, supported by the discovery in 1944 of DNA as the substance of heredity by Avery and his team (Beurton 2010). Finally, in 1953 Watson and Crick built a model for the structure of the DNA molecule and its replication, thanks to X-ray diffraction images of DNA taken by Gosling, Franklin and Wilkins (Clark and Pazdernik 2012). The molecular basis of the gene was found. From this moment, the history of the gene should be considered two-fold. On the one hand, the gene remained the so-called Mendelian gene, the unit of trait transmission and the object of classical genetics; on the other hand, the gene became identifiable with a precise chemical molecule (Kitcher 1984). The relations between the gene as the unit of transmission and the molecular gene are complex, and I will not enter into the details here. Suffice to say that Mendelian genes are not easily reducible to the molecular genes and they can be considered as something distinct with different functions and properties (Kitcher 1984; Okasha 2019). Here, I will not consider the Mendelian gene, and I will focus only on the molecular gene, the object of molecular genetics (Waters 2007).

[1] Three papers were published in the *Proceedings of the German Botanical Society* in 1900, by de Vries, Correns and von Tschermak (Fox Keller 2000). For the role of Mendel's 1866 paper in the birth of genetics, see El-Hani 2015.

In the 1960s, it seemed clear that genes were nothing more than segments of DNA located on a chromosome that gave rise to a particular amino acid sequence. The search for a material basis for the gene led scientists to postulate a correspondence between genes, segments of DNA and amino acids. This correspondence was formulated as the *Crick sequence hypothesis*: each codon, sequence of three bases, specifies only one amino acid, and a gene is a sequence of codons that specify for a polypeptide. The "material molecular gene" was born. Soon, a molecular understanding of the gene was accompanied by the idea that they were just open reading frameworks: DNA frameworks open to be read (ORFs). This facilitated research at the time, as the gene was identified as a well-defined and structured stretch of DNA, with clear borders and a singular function. The success of molecular genetics was read in terms of eliminativist reductionism. In 1969, Schaffner proposed to apply the models for epistemic reduction used in physical sciences to the case of the gene. If an identity relation between genes and DNA molecules had been found then, an eliminativist reduction would have been accomplished. The conclusions of this reductionist interpretation were more than epistemic, as they aimed at the *elimination* of the gene, and there was empirical and experimental support. A gene was claimed to be identical to a precise and defined stretch of DNA, gene1 = DNA1, as a simple 1:1 co-linear eliminativist reductive relation. "In the light of the Watson-Crick model, Benzer considered the possibility of translating his biological genetics into chemical terms" (Schaffner 1969, 339), and Schaffner regarded the entire reduction of molecular biology to chemistry and physics as something not only possible but very close in time.

However, things turned out to be more complicated than what Schaffner and many biologists thought in the 1970s. The production of new technologies to sequence genomes and the advances in molecular biology of the last 40 years have disclosed peculiar genetic phenomena (Fox Keller 2000; Hall 2001; El-Hani 2007; Griffiths and Stotz 2013; Meyer et al. 2013). The sequence of entire genomes and the study of eukaryotic genomes revealed that the Crick sequence hypothesis was simplistic, and further developments of genetics have made it impossible for genes to be a *merely* contiguous DNA segment co-linear with the product derived (Fogle 2010; Perini 2011; Rheinberger et al. 2015). These results compromised the material identity of the gene as a discrete stretch of DNA and showed the inefficiency of its identification in mere material terms (Falk 2010).[2] They also made clear how the gene operates within a system of complex and different mechanisms and processes. Particularly relevant to the beginning of the new phase of molecular biology is 2001 and the publication of the draft human genome sequence. This can be considered a "threshold year" as, from this point onwards, genetics entered the "postgenomic era" (as in Griffiths and Stotz 2013). In 2007, El-Hani spoke of the crisis of the gene that

[2]This does not imply that a given gene can never be considered within actual scientific practice as a precise and contiguous stretch of nucleotide sequences, such as in prokaryotes. Nevertheless, a strict identity relation between genes and genomic stretches excludes many genetic phenomena (see Fogle 2010; Griffiths and Stotz 2013).

finds itself between "the cross and the sword" because of the identification of a series of complex phenomena, such as split genes, alternative splicing, overlapping and nested genes.

2.1 The Gene in the "Postgenomic" World

Various attempts have been made to (re)-define the gene concept in the "postgenomic" world,[3] trying to accommodate both practical and theoretical requirements (as Beurton 2010). Generalising, we can point out two main ways to rethink the gene concept within the molecular context (Bellazzi 2022). The first is a deflationary instrumentalist approach that allows to retain gene speech and use without any further ontological commitment. The second is a realist approach that tries to define the gene embracing its complexity and context dependency. While both these approaches are informative about gene individuation, i.e. how to identify individual genes, here we are concerned with the gene definition or gene characterisation, i.e. what does it take for a given entity to be a gene.[4]

The deflationary approach to the gene is often referred to as "the nominal gene" and identifies it in an operational way on the basis of actual scientific practice and conditioned to research needs. According to this account, a gene can be *any* stretch of precise nucleotide sequences that encode a specific product (Burian 2004; Griffiths and Stotz 2006, 2013, 66). This approach leads to what is called the "consensus gene", as whatever pattern of "biochemical architecture and process" that presents the features of the exemplary gene, according to empirical evidence and scientific pattern (Fogle 2010; Bellazzi 2022). This approach to the gene remains nominalist or deflationary because it does not commit to the existence of the gene as something sui generis nor ontologically special: genes are any stretches of DNA that we find *useful* to identify as such given a particular model of gene. Moreover the identification of the gene remains conditional on research interests, maintaining minimal commitment to the entity.

In contrast to this approach, we also find a realist one, presented in the context of the ENCODE analysis.[5] Within this project, Gerstein and colleagues take the *gene to be an existent entity (and so independent from our research interests)* and define it as

[3] Often a pluralist view of genes is sustained, for which there are a variety of valid gene concepts in various disciplines (as Hall 2001, Fogle 2010).

[4] Havstad (2021) defines three classificatory practices concerning kinds: classificatory characterization or definition, individuation and organisation (2021). The first—which is the one we are concerned with—focuses on the definition of the kind. The second focuses on identifying which tokens belong to a given kind. The last focuses on organizing taxonomies.

[5] The ENCODE project is a project with the aim to identify all the "functional elements" in the human genome sequence. It represents an important project for the gene concept as it elucidated some complex phenomena that compromised the simple identification of the genes with contiguous stretches of DNA. Reference: https://www.encodeproject.org/about/2012-integrative-analysis

"a union of genomic sequences encoding a coherent set of potentially overlapping functional product" (Gerstain et al. 2007). This has been further re-elaborated by Griffiths and Stotz, who understand the "postgenomic genes" as existent "images of the target produced molecules" (Griffiths and Stotz 2013, 75). While this latter is not a definition,[6] it represents a helpful metaphor to understand the gene: it should present a non-necessarily contiguous sequence that is similar enough to the one of the transcribed molecule. According to this realist view, the gene is not simply identical to a linear and contiguous sequence, but a union of different ones. This union normally includes the finally transcribed sequence and the promoter region (or TATA box) (Fogle 2010, 6). In addition to it, the gene often comprises those regions essential for its activation and the regulation of its transcription, and these might be contiguous or not (Griffiths and Stotz 2013).[7] Following this realist approach in the formulation presented by Bellazzi (2022), a gene is an entity composed of those parts of the relevant nucleic acids that are transcribable (or involved in transcription) and encode a given mRNA. This second aspect, being transcribable, represents a functional component in the definition as the gene is a union of sequences with the function of encoding a target molecule. While the gene has a material component, according to this approach it is not only characterisable materially or as a material entity. The functional characterisation of the gene also allows us to embrace the context dependency pointed out by the postgenomic analysis: "a function is always a role in something and a contribution to something" (Germain et al. 2014). Transcription is not a self-subsistent phenomenon, but it is rather a reactive one: it happens only within the right circumstances and thanks to a set of interactions that operate at different levels. Accordingly, the gene can be fully understood only within such a context of action and interactions, and it is an entity defined materially and functionally (Falk 2010; Bellazzi 2022).[8]

In this paper, I will start from the second approach, in which the gene category refers to an existent union of sequences that are transcribable (or involved in transcription) and that transcribe precise genomic products (Bellazzi 2022).[9] This definition is coherent with contemporary genomics that is able to identify some

[6] I thank reviewer 1 for the helpful comments on this section.

[7] We can identify four main models that aim at identifying the precise material basis of genes. Model A presents the material basis of the gene as the transcribed region of the DNA plus all neighboring sequences which play a role in the process; Model B considers only the transcribed region, with introns and exons; Model C further restricts the basis and includes only the set of exons derived from a pre-mRNA; lastly, Model D limits the gene to only the coding exons of a primary transcript (Fogle 1990, El-Hani 2007, 3). Here, I leave open the question about the exact material component of the gene as it should be determined by scientific practice.

[8] As will be further clarified in Sect. 4.1, the functional component of the gene definition allows for both multiple realisation and multiple composition (in contrast with a materialist only view). A gene would be any entity that is composed of the relevant material aspect, nucleic acids—either DNA or RNA—and that has the relevant function. However, for individual genes the correspondence does not need to be 1 stretch: 1 function, as the function could be multiply realised by any stretch that realises it. For further references, see Bellazzi 2022.

[9] These sequences can either be of DNA or RNA according to the genes considered.

(often not contiguous) unions of sequences that take part in the transcription of given molecules (either amino acid chains or RNA molecules). But is the gene a *natural* kind? Or is it just a convenient category that groups some existent phenomena for practical need? In order to answer these questions, we should explore what it means to say that a natural kind is *natural*. In the next section, I am going to present an overview of the topic and the relevant account of natural kinds.

3 Natural Kinds and Biological Kinds

Division and classification of things into "sorts" or kind categories is common practice in both science and daily life. I recognise that *bread* is my favourite breakfast and that biologists study *proteins* and *amino acids* and find *microscopes* in their laboratories. We notice that some properties or features are co-occurrent in some individuals, and we cluster such groups of properties into kinds. These groupings are then associated with certain labels or predicates that allow the classification of individuals or relevant phenomena (Khalidi 2010). They allow us to make useful generalisations, being in our explanations and inductive inferences, and they proliferate from every corner of our life. Moreover, kinds play an especially important role in science. At least a part of scientific practice is based on clustering individuals into different categories and on making explanations and predictions about them (Bird and Tobin 2022). For instance, a biologist is generally not interested in the individual instance of a protein she is studying, or in the individual amino-acids string in her lab, but rather she aims at knowing something about the general *category* "protein". This would allow her to make generalisations valid across different instances. Moreover, the identification of these categories should provide some explanatory power and is often taken to be informative about the world.

Nevertheless, the ubiquity and variety of kinds brought the status of these categories into question. Philosophers have started to ask first what these categories are and second which of them can be deemed to be *natural* and which instead instrumental or artificial (Khalidi 1998, 2013). In order to explore the topic properly, it is important to distinguish two different enterprises, one concerning the kinds themselves and one concerning the naturalness of the categories that refer to them. More precisely, we can summarise two main questions that reflect the discussion on natural kinds (Magnus 2015, 2018):

- An *ontological question* that asks which kind of entity, if it is an entity at all, a natural kind is. And an answer might come from a theory of sui generis universals, cluster of properties or similarity between instances.[10]

[10] For an overview on the debate on universals, consider Bird and Tobin 2022.

- A *naturalness question*[11] that asks how we can recognise an arbitrary category from one that captures some genuine divisions in nature. What must a category do in order to "carve nature at its joints"?

In this paper, we are inquiring into whether genes are something more than a useful tool to do genomics, and we will focus on the second question in order to do so. This allows us to explore the status of the category without considering universals, essences or questions within fundamental ontology[12] (Magnus 2018). We can identify two broad strategies to answer the natural question. A first is offered by conventionalism: there are no natural kinds, but only conventions that suit different purposes. A second is offered by a form of realism: at least some of our categories correspond to natural kinds and the objective features of reality. It is common to take a hybrid position. Some kinds that we find in scientific practice or daily life are conventions, while others might be correspondent to real features of the world. For the sake of the present analysis, I will assume a form of minimal realism for which at least some categories correspond to genuine features of reality and might be candidates for natural kinds (Khalidi 2013; Bird and Tobin 2022). However, not all the categories that we identify as kinds seem to correspond to such divisions. How do we distinguish "natural kinds" from mere "human categories" though (Khalidi 2013)?

In the biological sciences, this question is particularly relevant as the study of life appears to be the reign of taxonomies and classifications. Historically, species have been taken to be a paradigmatic case of kinds, and the concerns on whether all of the *Linnean taxa* correspond to actual divisions in reality have been widely debated.[13] In order to identify natural categories, some accounts have been proposed, among which are essentialism in intrinsic and historical forms, HPC cluster theories and others. However, most of them are concerned with the question of whether any of the *taxa* can be considered a natural kind (Slater 2013). This makes them interesting for the species or higher taxonomies debate but makes them less applicable to other candidates of natural kinds in biology, such as kinds in biochemistry or at other levels (Slater 2013; Khalidi 2013; Kistler 2018). In this paper, we are exploring a category that comes from the domain of genetics, and we are asking whether unions of genomic sequences identified as a gene can be considered instantiations of a natural kind or not. Thus, in order to explore an answer to this case of the naturalness question, let me present what it takes for a category to be natural.

[11] This question can also be referred to as the "taxonomy question". However, I prefer to avoid this terminology as it seems to constrain natural kinds to biological taxonomies or to identify kinds with taxonomical classifications.

[12] See Magnus' definition of *deep realism* (2018) or Khalidi's definition of *Realism* (2013) for which one has to commit to the existence of some fundamental categories.

[13] A summary of the status of the species controversy can be found in Ereshefsky (2017).

3.1 Natural Kinds

The debate on natural kinds is wide in approaches and topics. Nevertheless, there is a consensus on the fact that kinds should be those that allow to make reliable explanations and predictions across instances (Bird and Tobin 2022). This feature is represented and clearly discussed, together with other accounts, in Khalidi's approach (1993, 1998, 2013, 2018, 2020, 2021). His view presents kinds as those categories present in scientific theories that are projectible and capture nodes in causal networks of the world. This approach stands out because it is a realist account, for which kinds track objective features of reality while at the same time avoiding excessive metaphysical commitments. Moreover, it is applied and applicable to a variety of kinds from the physical and special sciences, being able to take into account both structural and historical or etiological properties.[14] Accordingly, I take his view as a starting point to explore whether the gene could be deemed a natural kind. While my argument is conditional upon such account, I do not think that this compromises its validity as this account has commitments on projectibility and causal efficacy, which are often considered valid criteria for natural kinds (as in Bird and Tobin 2022). Let me present Khalidi's account in more detail.

Khalidi is interested in exploring a way to answer the natural question and to identify and distinguish conventional groupings from objectively existent ones. He starts off with a form of *weak or moderate realism (r)* for which kinds are objective features of reality, but not corresponding to distinct metaphysical categories.[15] This means that nature has some joints as objective features of the world (whether we know them or not), and our best theorising should aim at carving up these joints. However, not all of our categories are natural kinds, and some can be considered mere conventional grouping. His account aims at providing a way to identify *natural* kinds. To do so, he claims that we should start looking at our best science and scientific practice within both natural and social sciences (2013, 2021). Specifically, he sustains a form of moderate naturalism, for which natural kinds can be taken to be some of the categories revealed by our systematic attempts to gain knowledge of nature. Taking a realist stance towards the discipline, science aims at identifying kinds that are really existent in the world, and not mere and useful theorisations, and it has proved so far rather successful in doing so. So, if we want to disentangle and identify the *natural* kinds among all the categories, then a good starting point is to look at different sciences. Moreover, the combination of a weak form of realism and a form of naturalism allows us to clarify an important aspect of the theory of natural kinds: science or the philosophy of natural kinds does not invent natural kinds, but

[14] Khalidi's account has also been applied to a variety of kinds from the life sciences, such as viruses, cancer cells, biological species and ADHD, making this approach even more suitable to consider the gene case (Khalidi 2013, 2021).The validity of Khalidi's view of natural kinds is discussed also in Tahko (2022).

[15] This view is contrasted with a stronger form of *Realism* in which kinds correspond either to *sui generis universals* or to *second-order universals*.

rather discovers them. Scientific theories or established knowledge does not determine the existence or the non-existence of kinds, but rather they represent our *access and guide* to the existence of such kinds (Khalidi 2013).

Nevertheless, even within the best scientific theories, the history and philosophy of science have shown that not all categories present in the discipline can be considered as capturing something in the world. For instance, a category like *hysteria*, which has been used as a scientific category in the past, has proved not to be a natural kind of diseases and was abandoned as a kind (Khalidi 2013, 59). Moreover, some categories can have an instrumental role or cannot be considered stable or robust enough to be really informative about the world. This led Khalidi to add two further requirements for naturalness.[16] These are (i) genuine projectibility and (ii) being a node in a causal network.

Projectibility is often assumed as a feature that natural kinds should display. Kinds are particularly efficacious categories when it comes to framing inductive inferences, and they feature in many empirically verifiable generalisations. This means that kinds are projectible, in the sense that they can be projected from one instance to another in a successful way. More precisely, the projectibility of a natural kind can be defined as follows: "when it comes to a natural kind predicate K, there is no shortage of other predicates, P1, P2, . . ., Pn, and so on, such that we can reliably assert that if x is K, then x is P1, x is P2, . . ., x is Pn and we can do so with a high degree of generality" (Khalidi 2018, 1385). Kinds provide explanatory and predictive power across different contexts and circumstances because they allow us to project a set of properties from one instance to the other and to predict that such properties will be present (Khalidi 2013, 2018, 2020, 2021).

Moreover, this kind of projectibility requires an explanation: why is it possible to draw these inferences? What is the ontological ground for which the kind can be applied in an explanatory way to many instances? Khalidi answers these questions by adding a second requirement that natural kinds should satisfy: they are "nodes" in causal networks. Projectibility results as a "reflection" of the causal network in which *instances* of the kinds are involved, and some kind categories are particularly successful because the properties of the natural kinds are *causally clustered* (Khalidi 2013, 2018, 2020).[17] The joints that natural kinds carve so successfully are those that can be found in the causal structure of the world. Together with providing an ontological reason for the projectibility of kinds, causal relations also provide an answer to the naturalness question. Specifically, they play two main roles in distinguishing natural from conventional kinds. First, natural kinds do not only present a set of projectible properties but a set of properties that are hierarchically ordered as "causes and effects in recurrent causal processes" (Khalidi 2018). They

[16] These are based on a re-elaboration of the simple theory of natural kinds proposed by Craver (2009) for which kinds refer to the causal structure of the world.

[17] It is important to notice that the natural kinds display a role in the causal network which can be seen when considering instances of the kinds due to the nature of the causal relation. I thank Jessica Wilson for suggesting this important clarification and reviewer 2 for insisting on this aspect.

present a set of "core" properties that cause the instantiation of other properties of the instance of a given kind. Natural kinds are those categories with a set of properties discoverable by science and whose co-instantiation *causes* the instantiation of other properties (Khalidi 2013). Second, natural kinds are those categories that represent nodes within broader causal processes: they are causally efficacious on other kinds and are intersections within the webs of causal relations. The causal relations among property instances and the causal cores of natural kinds represent the ontological principle in virtue of which we can distinguish natural from unnatural kinds. Moreover, it is the underpinning of the projectibility of such categories.

To further elucidate the account, let me consider briefly an instance of a natural kind presented by Khalidi: the case of viruses (2013, 180). Viruses represent an established category of a proper subdiscipline, virology. Virions, individual particles of the virus, are characterised by the identification of some synchronic causal properties. First, they are protein particles (more or less complex) containing a genome capable of making an mRNA readable by the ribosome of a host cell. Second, the virions of a given virus display a specific infectious cycle comprised of well-understood causal relations. Following Khalidi's methodology, we need to consider whether viruses present the three main features of naturalness: (i) being present in scientific theories, (ii) being projectible and (iii) being nodes in causal networks. The first criterion is easily satisfied, as our analysis started with a scientific category present in an established enough discipline. Let me consider the other two. The infection cycle and the life cycle of viruses are understood in terms of common and repeatable causal processes that display a given hierarchy of relations. These causal processes allow virologists to make empirical generalisations on the different instances of viruses and their cycle. Such generalisations can then be projected from viruses already observed to the ones that have not been. This projectibility is broad and stable across contexts: different kinds of viruses in different circumstances present a cycle that is re-conducible to the general causal one identified by virology. This makes virus a projectible kind: it allows explanatory and predictive power across different instances of the kind. Furthermore, projectibility is sustained by the fact that viruses can be taken to be appropriate nodes in causal networks. First, the core properties of the virus, such as being a protein containing a genome capable of making an mRNA readable by a host cell, *cause* the instantiation of other properties, such as a given infection rate and behaviour within the host cell. This orders the properties that can be ascribed to the virus category according to a causal hierarchy. Second, viruses enter into causal interactions in a uniform or similar pattern, and they have a causal impact on the network of relations in which they are embedded. Viruses are causally efficacious categories within the network of relations they are involved in. Accordingly, the category virus is a natural kind as it is the object of a successful part of contemporary science, it is a projectible category and such projectibility is based on causal relations. Specifically, the properties that pertain to the kind as a whole are causally related to each other and lead the members to enter into causal interactions in a uniform way.

The theory presents some advantages compared to the ones previously analysed. First, this account is a good example of the consensus reached by philosophers on

how to answer the naturalness question, providing us with a way to disentangle natural from unnatural kinds. This is done without embracing an essentialist theory that tries to provide a specific set of properties for which an individual is a member of a kind. Moreover, it is an application of a form of reflective equilibrium between scientific input and philosophy. It is a combination of convictions on categories generally regarded as paradigmatic kinds, often taken to be stable categories in scientific theories, philosophical discussions on natural kinds and a set of considerations that are drawn from scientific practice. Lastly, this account can be applied to a variety of different types of kinds. It is able to accept etiological and historical kinds as natural kinds, considering a particular origin or genealogical history as the core properties that cause the instantiation of other ones. Moreover, the combination of naturalism with projectibility and causality allows the theory to be applied to concrete case studies within fundamental physics and the special sciences, such as lithium, cancer cells, viruses and ADHD (examples from Khalidi 2013).

Concluding, this account answers the naturalness question and will be used in the next section to explore whether genes can be considered a natural kind.

4 The Gene as a Natural Kind

The naturalness of the gene category has been questioned because of its aforementioned history, the complexity of the genetic phenomena and the context dependency that is implied by genes' functional aspect. Genes might not seem neat enough to count as a natural category or might be too ambiguous. In this section, I am going to argue for the opposite, defending a view for which the gene can be deemed a natural kind. Before pursing my thesis, some clarifications are in order. This is because the genomic and biochemical domain is rich in systems of practice and taxonomies, and clarity is particularly needed to avoid ambiguities. According to Havstad (2016, 2021), we need to consider three classificatory practices concerning kinds: classificatory characterisation or definition, individuation and organisation. The first focuses on the definition of the kind, already presented in Sect. 2.1. The second focuses on identifying which tokens belong to a given kind. The last focuses on organising taxonomies. Accordingly, before assessing the naturalness status of the gene category, I will consider briefly gene classifications and taxonomies. Then, in Sect. 4.2, I will explore whether the gene can be considered a natural kind.

4.1 Gene Classifications and Taxonomies

As presented in Sect. 2.1, the gene is defined in the context of the ENCODE analysis as "a union of genomic sequences encoding a coherent set of potentially overlapping functional product" (Gerstain et al. 2007). According to this approach, genes are those (not necessarily contiguous) sequences that have a specific function in

encoding a given molecular product. The classification of tokens or instances of genes into types or families is mostly done according to the product: unions of sequences encoding the same product (or products that are similar enough) are clustered into the same type of genes.[18] The specific function of a gene within transcription is what allows to classify them.[19] Moreover, such classification of genes is done at different levels. First, there is a broad classification of genes into two functional subtypes: (i) genes that encode regulatory RNAs that play different functions within cellular processes and (ii) genes that encode an RNA for the amino acid sequence of a polypeptide (Perini 2011). Then, we can find more specific classifications of gene tokens into gene types according to the given molecular product of the gene under consideration. Tokens of the same type of genes are clustered together if they encode the same molecular products or molecular products that are considered similar enough (Gerstain et al. 2007; Fogle 2010; Griffiths and Stotz 2013). In line with the importance of the link between genes and products, protein-coding genes (those encoding an RNA for a polypeptide) are normally named after the protein they encode. In this classificatory system, the *locus* is indicated often in italic capital letters while the name of the gene in capital letters and the protein made in normal characters. An instance of gene classification based on DNA sequences and function is the one of the genome of *Saccharomyces cerevisiae*, sequenced in 1996 and constituted of about 6275 genes, organised on 16 chromosomes. Of these genes, about 5800 are identified by their function.[20] For example, the gene DCS2 encodes the protein dcs2, which takes part in the biological processes that regulate the response of the cell to heat (Liu et al. 2017). Generalising, the link between genes and products, together with the consideration of the relevant union of genomic sequences, is what normally allows gene individuation and gene talk.

As far as other taxonomies or nomenclatures are considered, genes are classified in a variety of ways often referred to as gene ontologies. In order to aid scientific practice, these ways of classifying the genes have been grouped within the project *The Gene Ontology*, a resource that provides a computational representation of our current scientific knowledge about the functions of genes according to the functions of the product they encode. Overall, this project considers three aspects when classifying genes: (i) molecular functions performed by gene products, (ii) cellular components in which the gene product performs a function and (iii) biological processes in which gene products are involved. For example, the gene for the product "cytochrome c" can be classified according to the molecular function of

[18] For the importance of functional similarity of products in gene classification, see Fogle 2010.

[19] Despite the importance of the functional aspects of the genes, genes are not mere functional kinds, because the material component they present as union of genomic sequences is also relevant in determining their identity.

[20] For the database on *S. cerevisiae* genome and the relevant articles, see www.yeastgenome.org

the product (oxidoreductase activity), the biological process (oxidative phosphorylation) and the cellular component (mitochondrial matrix) (Gene Ontology[21]).

Lastly, it is possible to find different ways to organise genes in taxonomies contingent on specific contexts and given scientific goals. For example, in evolutionary developmental biology and comparative genomics, genes are often classified in terms of evolutionary history. However, these taxonomies tend to be highly context relative and presuppose the aforementioned characterisation of the gene. First, scientific practice clusters tokens of genes across species into the same type according to a given union of sequences and a given function in encoding a product that takes part in molecular and cellular processes. Then, thanks to the similarity of the sequence and the RNA encoded, they can be further classified in terms of evolutionary similarities. An example is constituted by genes that can be classified as *homologous* if they are inherited by two different species from a common ancestor, and such classification is done in terms of sequences and RNA encoded.

To conclude, genes can be classified and organised in taxonomies according to different practices and in a variety of ways. Nevertheless, this classification presupposes an initial identification of the gene as a given union of genomic sequences that encodes an amino acid sequence. Once the gene is individuated, then it can be classified for further purposes. This is possible thanks to a general characterisation of the gene concept, as that union of genomic sequences that are transcribable in a target molecule (Bellazzi 2022). Such characterisation of the gene is the one relevant for the purpose of the paper.

4.2 *The Gene* Is *a Natural Kind*

Having summarised how genes can be organised in taxonomies, I will now move on to consider whether the category gene itself can be deemed a natural kind. To do so, I will start considering whether the properties associated with the natural kind category respect the requirements for naturalness presented in Sect. 3.1.

In Sect. 2, I have pointed out the concept of gene relevant for our analysis. Genes result from the union of sequences transcribable in a target molecule, and they play their function only in a wider system of interactions and environments (Griffiths and Stotz 2013). According to this definition, the gene presents a two-fold identity: in terms of structure or composition and in terms of function. First, the gene has a structural component as a region or union of regions of nucleic acids, while not every region of the given nucleic acid is a gene (Fogle 2010; Griffiths and Stotz 2013). Specifically, the gene is composed of those regions that can be considered images of the target molecule and are actively involved in transcription. This leads us to the second aspect of the gene. It is the union of sequences that has a given function in encoding the primary structure of a polypeptide or of a functional RNA molecule.

[21] For Gene Ontology, see www.geneontology.org

The functional component of the identity of the gene is evident during transcription, a reactive process that can be upregulated and downregulated thanks to specific intra-, inter- and extracellular interactions. This makes the gene a context-dependent entity and a multiply composable one (Germain et al. 2014; Bellazzi 2022). Specifically, a given instance of the gene kind can be composed of different stretches of nucleic acids, while the function characterising the gene is maintained (i.e. the final transcribed molecule).[22] This feature is important to account for the complexity of genetic phenomena. Accordingly, to understand what a gene is, one needs to consider, first, the identification of those sequences that are transcribable in a target molecule and, second, the process of transcription where the gene plays its function. The identification of the main features of the gene category is the starting point in the analysis of whether the gene can be considered a natural kind.

Now, we need to explore whether these properties respect the three main requirements for naturalness: (i) naturalism, the presence of this category in our best scientific theory; (ii) projectibility, whether the category has explanatory power and can be projected from one instance to another; and (iii) whether the gene is a node within a causal network.

Let me start with naturalism. Despite their complicated history and the current challenges raised by postgenomic analysis, genes still retain an important role across a different variety of life sciences. Specifically, they are often invoked in the study of protein synthesis and the impact that alterations in protein structure can play at the cellular and organismal levels. Moreover, genes still retain a role in developmental and evolutionary biology. The gene is a category that figures across different scientific disciplines in a stable way, thus meeting the first requirement for naturalness. Genes also present projectibility and being nodes in causal networks. First, we can project the properties of the genes from one instance to another in a successful way. This is particularly evident in protein synthesis, one of the phenomena in which genes play an important role. In this case, scientists project the two main properties of the genes from one instance of protein synthesis to another in order to be able to explain and predict general patterns. At the same time, protein synthesis shows how genes are nodes in causal network: they have a causal role in the synthesis of proteins, and the two properties are causally related for this to happen.

To further support my argument and appreciate why genes respect the natural kinds requirements, I will consider a specific case study: the aforementioned gene DCS2 in *Saccharomyces cerevisiae*. This gene encodes the protein dcs2, which takes part in the biological processes that regulate the response of the cell to heat (among other processes) (Liu et al. 2017). Furthermore it presents clearly the two properties of the category "gene" (being composed of nucleic acids and having a function),

[22] In order to appreciate how the definition allows for multiple realisability and composition, it is important to consider that the gene *type* is the one that can be multiply composed, while each gene *token* is going to be composed by specific nucleic acid sequences. Accordingly, for the gene definition we just need to identify which specific token nucleic acids could compose it while maintaining that the relevant function is realised. For further analysis of the complexity of the relations between the properties of the gene, see Bellazzi 2022.

making it a good case study to see how these properties support the naturalness of the kind.[23]

Following the characterisation of the gene presented here, this gene is transcribable in a given target molecule, the mRNA for the protein dcs2, and (i) it has a specific basis on chromosome XV of the cell and (ii) has a specific function, i.e. encoding the protein dcs2.

Let me now then consider if this gene is projectible, that is, if it allows better explanations and predictions based on the causal properties it presents. The gene DCS2 encodes the mRNA for the protein dcs2, a regulatory protein that acts as a pyrophosphatase regulator in many cellular processes and specifically in the heat response of the cell.[24] The instances of this gene have a common (type) nucleic acid base, which makes the relevant gene identifiable on a precise locus on chromosome XV of the cell, and they encode a given protein within the right circumstance. Also, instances of this gene take part in the process of transcription which leads to the presence of the protein within the cell, thereby taking part in a causal process. The same role is played by distinct instances of this gene across different cell individuals, and the identification of this causal pattern allows to better predict and explain heat regulation processes. Moreover, this gene can be deemed homologous to the gene DCS2 in other species such as *Homo sapiens* or *Mus musculus* allowing this category to be projectible not only across instances of *Saccharomyces cerevisiae* but also to other species. Accordingly, the gene DCS2 is projectible in virtue of having the properties of the kind "gene". This example illustrates how it is possible to project the category "gene" from one instance to another in order to make reliable explanations and predictions about genetic phenomena.

Let me consider now whether genes can be considered nodes in causal networks. First, we need to see if the core properties of the gene can be considered causally related. Second, we need to consider whether the gene plays a causal role within a broader context. I will do so with the help of the previous example. The two properties of the gene DCS2 are (i) being composed of a given union of sequences and (ii) encoding for the protein DCS2. The two seem to be at least partially causally related as the sequence allows for the interactions necessary for the transcription of the DNA sequence into the mRNA that then encodes the protein dcs2. The "cross-talk" between the sequence, RNA polymerases and the actors of transcription is what causes the second definitional property of the gene to be present. The sequence has a causal role in the manifestation of the functional property together with the relevant factors, allowing the encoding for a specific protein. This supports the presence of a causal link between the two main properties of the gene and allows a hierarchical causal characterisation of the other properties that might be an effect of those two.

[23] The purpose of this section is not to show that "DCS2" is a gene, but rather that the properties that genes display (and are present in DCS2) respect the requirements for the naturalness of kinds. Moreover, the causal component required by the naturalness claim supports using an instance of the kind to explore whether it is natural. I thank reviewer 2 for insisting on this clarificatory point.

[24] Further reference can be found at *yeastgenome.com*.

Furthermore, the gene under consideration plays a role within the general causal network. For instance, the encoding of the protein dcs2 contributes causally to the heat regulation within the cell. This can be generalised even further given that genes can take part in a variety of causal phenomena. First, and most evidently, as aforementioned, they represent a node in the causal network that is identifiable in protein synthesis. Second, contemporary molecular biology has found out that genes can encode a variety of RNA molecules that play different roles. This makes the gene a node within a series of cellular processes. As follows, the gene also respects the second requirement: they have causally related properties that, within the right systems of interactions, can cause further phenomena. Lastly, having considered the features of a specific gene to support the argument does not compromise the generality of the conclusion because the discussed properties of the gene DCS2 are those characteristic of the kind "gene"—being composed of nucleic acids and being transcribable in a target molecule—and not those specific of DCS2. Accordingly, similar considerations can be made for other instances—even more complex ones—in virtue of the two core properties of the kind.[25]

In conclusion, the gene category captures a natural kind according to the requirements for the naturalness of a category. First, this category has an important role within our scientific theories, respecting a naturalist approach. Second, it is possible to identify two main properties that are important for gene definition. These properties allow its identification, and that can be used as a basis for the analysis of naturalness. The gene category is projectible, as its definitional properties have a predictive and explanatory role, and scientists can project results about one instance to other instances on the base of its core properties. Moreover, these properties can be deemed to be causally related to each other, and the gene can be considered a node in a causal network in different processes.

5 Conclusions

In this paper, I have argued that the gene can still be retained as a natural kind within the postgenomic context, once we accept Khalidi's view of natural kinds. In the first part, I have briefly illustrated the history of the gene and why it might be considered a controversial category. The discoveries that followed the ENCODE project have shown that it is difficult to identify the gene with a precise stretch of DNA and that it is important to consider a full system of interactions to understand the complexity of

[25] I thank both reviewer 1 and reviewer 2 for the suggestion to clarify the relation between the example of gene considered in the paper and the general kind "gene". The definition of gene defended in the paper allows the generalisation from simple to more complex case studies, as the definition presented here is compatible with both multiple realisation and composition. Accordingly, we can generalise it to cases such as those involving alternative splicing because it is possible to identify the two core properties of the kind "gene" also in those cases (see Bellazzi 2022 for more discussion on this).

genetic phenomena. A possible reaction to such controversy is to take a nominalist view of the gene and consider it as a conventional or instrumental category. However, the success of the genes category across different sciences and its wide applicability can be considered a hint in support of the naturalness of the kind.

In order to argue in favour of the genes as a natural kind, I have illustrated in Sect. 3 what makes a given category *natural*. To do so, I have used Khalidi's account as a starting point for the analysis. Specifically, a kind is said to be natural when it respects three requirements: (a) it figures in our scientific theories; (b) it is projectible; (c) it is a node in the causal network. Then, in Sect. 4, I have illustrated why the gene is a natural kind. This has also bought in necessary considerations on genes taxonomies and classifications into genes type and tokens. The definitional properties of the gene are individuated mostly synchronically and respect the criteria proposed for the naturalness of the kind. First, the gene category figures in scientific theories, respecting the naturalist requirement. Then, the individuated properties of the gene allow the members of the kind to be part of generalisations and projections concerning the category. This is possible because the properties are causally related to each other and allow the instances of the kind to enter into causal interactions that are uniform and identifiable.

Concluding, while the argument is conditional upon the acceptance of Khalidi's view, a realist approach to genes as natural kinds has some benefits. First, if we accept that one of the aims of science is to discover what kinds of things are in the world, then an argument in favour of the naturalness of the gene category supports the success of genetics. Second, having a realist or instrumentalist view towards a scientific category might have an impact on scientific practice and scientific discoveries. An example of this can be found in the history of the gene itself, summarised at the beginning of this paper. The realist understanding of the gene supported by Muller offered a theoretical framework for which scientists started looking for the gene as a material entity rather than a mere theoretical instrument. This contributed to the discovery and identification of the gene as specific stretches of DNA or as at least located on the DNA. A realist understanding of the gene category seemed to have had an impact on the direction of research. Accordingly, the consideration of the gene as a natural kind can have an impact on how scientists think about this category, even if they might not change scientific practice on a daily basis. Lastly, conceiving the gene a category that "carve nature at its joints" rather than a mere instrument might bring clarity to the debate and offer a broader framework for future research.

Acknowledgements I would like to thank Tuomas Tahko, Samir Okasha and Elle Chilton-Knight for reading and reviewing multiple versions of this paper. For the helpful and constructive comments that have substantially improved the paper, I would like to thank the anonymous reviewers. I am sincerely grateful to Maurizio Zuccotti and his team for the insightful discussions on the genetics case studies. Thanks also to the members of the ERC MetaScience Project – Toby Friend, Vanessa Seifert, Samuel Kimpton-Nye – and to Jessica Wilson, Giacomo Zanotti, Margarida Heremida and Luca Zanetti for the helpful discussion, advices and feedback. Lastly, I would like to thank deeply the editors and the organisers of the PBCS X workshop and the audiences where I have presented the main ideas of the paper.

References

Bellazzi F (2022) The emergence of the postgenomic gene. Eur J Philos Sci 12:17

Beurton PE (2010) A unified view of the gene, or how to overcome reductionism. In: Beurton PJ, Falk R, Rheinberger HJ (eds) The concept of the gene in development and evolution. Cambridge University Press, Cambridge, pp 286–314

Bird A, Tobin E (2022) Natural kinds. In: Edward Zalta N, Nodelman U (eds) Stanford Encyclopedia of philosophy. Spring. (2023 Edition). https://plato.stanford.edu/entries/natural-kinds/

Burian RM (2004) Molecular epigenesis, molecular pleiotropy, and molecular gene definitions. History and Philosophy of the Life Sciences 26 (1-Genes, Genomes, and Genetic Elements):59-80

Clark D, Pazdernik N (2012) Molecular biology. Elsevier Science Technology, Academic Cell

Craver C (2009) Mechanisms and natural kinds. Philos Psychol 22(5):575–594

El-Hani CN (2007) Between the cross and the sword: the crisis of the gene concept. Genet Mol Biol 30(2):297–307

El-Hani CN (2015) Mendel in genetics teaching: some contributions from history of science and articles for teachers. Sci & Educ 24:173–204

Ereshefsky M (2017) Species. In: Edward Zalta N, Nodelman U (eds) Stanford Encyclopedia of philosophy. https://plato.stanford.edu/entries/species/

Falk R (2010) The gene—a concept in tension. In: Beurton PJ, Falk R, Rheinberger HJ (eds) The concept of the gene in development and evolution. Cambridge University Press, Cambridge, pp 317–348

Fogle T (1990) Are genes units of inheritance? Biol Philos 5:349–371

Fogle T (2010) The dissolution of protein coding genes in molecular biology. In: Beurton PJ, Falk R, Rheinberger HJ (eds) The concept of the gene in development and evolution. Cambridge University Press, Cambridge, pp 3–25

Fox Keller E (2000) The century of the gene. Harvard University Press, Cambridge MA

Germain PL, Ratti E, Boem F (2014) Junk or functional DNA? ENCODE and the function controversy. Biol Philos 29:807–831

Gerstain MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681

Griffiths P, Stotz K (2006) Genes in the postgenomic era. Theor Med Bioeth 27:499–521

Griffiths P, Stotz K (2013) Genetics and philosophy. Cambridge University Press, Cambridge

Hall BK (2001) The gene is not dead, merely orphaned and seeking a home. Evol Dev 3(4):225–228

Havstad JC (2016) Protein tokens, types, and taxa. In: Kendig C (ed) Natural kinds and classification in scientific practice. Routledge, New York

Havstad JC (2021) Complexity begets crosscutting, dooms hierarchy (another paper on natural kinds). Synthese 198:7665–7696

Khalidi MA (1993) Carving nature at the joints. Philos Sci 60(1):100–113

Khalidi MA (1998) Natural kinds and crosscutting categories. J Philos 95(1):33–50

Khalidi MA (2010) Interactive kinds. Br J Philos Sci 61(2):335–360

Khalidi MA (2013) Natural categories and human kinds: classification in the natural and social sciences. Cambridge University Press, Cambridge

Khalidi MA (2018) Natural kinds as nodes in causal networks. Synthese 195:1379–1396. https://doi.org/10.1007/s11229-015-0841-y

Khalidi MA (2020) Are sexes natural kinds. In: Dasgupta S, Weslake B (eds) Current controversies in philosophy of science. Routledge, New York

Khalidi MA (2021) Etiological kinds. Philos Sci 88:1–21

Kistler M (2018) Natural kinds, causal profile and multiple constitution. Metaphysica 19(1): 113–135

Kitcher P (1984) 1953 and all that. A tale of two sciences. Philos Rev 93(3):335–373

Liu W, Li L, Ye H, Chen H, Shen W, Zhong Y, Tian T, He H (2017) From Saccharomyces cerevisiae to human: the important gene co-expression modules. Biomed Rep 7:153–158
Magnus PD (2015) John Stuart mill on taxonomy and natural kinds. J Int Soc Hist Philos Sci 5(2): 269–280
Magnus PD (2018) Taxonomy, ontology, and natural kinds. Synthese 195:1427–1439
Meyer LMN, Bomfim GC, El-Hani CN (2013) How to understand the gene in the 21st century. Sci Educ 22(2):345–374
Okasha S (2019) Philosophy of biology: a very short introduction. Oxford University Press, Oxford
Perini L (2011) Sequence matters: genomic research and the gene concept. Philos Sci 78(5): 752–762
Rheinberger HS, Muller-Wille S, Meunier R (2015) Genes. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/gene/
Schaffner K (1969) The Watson-Crick model and reductionism. Br J Philos Sci 20:325–348
Slater MH (2013) Cell types as natural kinds. Biol Theory 7(2):170–179
Tahko T (2022) Natural kinds, mind-independence and unification principles. Synthese 200(2): 1–23
Waters K (2007) Molecular genetics. The Stanford Encyclopedia of Philosophy, Fall 2013 edn. Edward N. Zalta (ed). https://plato.stanford.edu/archives/fall2013/entries/molecular-genetics/

Part IV
Teleology in Biology and Cognitive Sciences

Teleological Explanations and Selective Mechanisms: Biological Teleology Beyond Natural Selection

Javier González de Prado and Cristian Saborido

Abstract From a naturalistic approach, several attempts have been made to justify teleological explanations by appealing to the action of selective mechanisms. In philosophy of biology, natural selection has often been assumed to be the paradigmatic case of selective mechanism, and, on this basis, different generalized biological selective explanations have been proposed in an attempt to substantiate natural teleology. In this paper we use a different strategy. Starting from a general definition of selection as differential reinforcement, we interpret the different types of teleological explanation, both biological and non-biological, as specific cases of selective explanations, of which evolutionary explanations would be only a specific subset (rather than the only ones). We illustrate this by analyzing teleological explanations that make reference to biological regulatory processes.

Keywords Teleology · Selection · Function · Evolution · Regulation

The authors are listed alphabetically. Both of them have contributed equally to this work. We are especially grateful for the feedback from the editors of this volume, as well as from the PBCX attendees. We also thank audiences at the University of Bielefeld, Complutense University (Madrid), the University of Helsinki, the National University of Colombia (Bogotá), Nova University of Lisbon, the University of Oslo, the University of Prague, UAM (Madrid), the University of the West of England Bristol, and UNED (Madrid). Special thanks to Leonardo Bich, Matteo Mossio and several anonymous reviewers. Thanks as well to Megan J. Watkins for the linguistic revision of the article. This work has been supported by the Spanish Government research projects APID2021-128835NB-I00 and PID2021-123938NB-I00.

J. González de Prado (✉) · C. Saborido
Department of Logic, History and Philosophy of Science, National Distance Education University, Madrid, Spain
e-mail: jgonzalezdeprado@fsof.uned.es; cristian.saborido@fsof.uned.es

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_13

1 Introduction

According to what is perhaps the most popular account of biological teleology, biological purposes are introduced by natural selection (Millikan 1984, 1989; Neander 1991; Griffiths 1993; Kitcher 1993; Godfrey-Smith 1993; Buller 1998, Artiga 2021). The function of a biological trait, in this type of view, is to do whatever previous tokens of that trait were selected for by natural selection. This account can be seen as an instance of selected-effects theories of teleology, which hold that purposes are effects for which an item has been selected.

Selected-effects theories, as we will see, offer an attractive account of many paradigmatic forms of teleology, such as the functions of artifacts, or goal-directed behavior in intentional, rational agents. Arguably, when there is selection, there is teleology. Thus, a promising way of vindicating biological teleology is to argue that there are biological processes, in particular evolution, that involve genuine cases of selection and then to appeal to selected-effects theories of teleology. This vindicatory strategy requires showing that the alleged biological selective processes share the distinctive features of paradigmatic types of selection and, therefore, should be considered as genuine forms of selection as well. This provides motivation to look for a generalized account of selection that covers both biological selection and paradigmatic instances of selection in other domains.

Further motivation for a generalized account of selection is that it opens the door to the recognition of new types of biological teleology. In principle, there could be biological processes other than natural selection that exhibit the defining characteristics of selective processes. If we have a generalized account of selection, which does not apply just to natural selection, we can check whether a given biological process deserves to be considered a form of selection.

When constructing a generalized account of selective processes, one approach is to take natural selection as our model. In this approach, a process counts as selective insofar as it is analogous or relevantly similar to natural selection. We want to argue, however, that this approach is on the wrong track. Natural selection differs in several significant points from many paradigmatic types of selection. Thus, modelling a generalized account of selective processes on natural selection has the risk of leading to an unduly restrictive account.

A potential negative consequence of this is that one may fail to count as forms of selection biological processes that, despite not fitting the mold of natural selection, closely resemble paradigmatic types of selection. We think that this is what happens, for instance, with biological regulation. As we will explain, biological regulation shows relevant dissimilarities with natural selection. One could claim that, due to this, it is a stretch to regard regulation as a type of selection. However, we will argue that, if anything, regulation is closer than natural selection to paradigmatic forms of selection.

Our claim, therefore, is that a lack of similarity to natural selection is not a reason to discard processes that are strongly analogous to more paradigmatic types of selection as nonselective. As a kind of selection that in important ways differs

from many paradigmatic selective processes, natural selection should not be taken as the (only) yardstick to determine what counts as selection. Instead, we propose constructing a minimal, generalized account of selection that captures both paradigmatic forms of selection and more atypical cases. We suggest doing so by relying on the notion of differential reinforcement. Biological regulation, as we will see, fits this generalized account and, therefore, deserves to be considered a genuine type of selection—giving rise, according to selected-effects theories, to its own form of teleology.

2 Selected-Effects Theories

Teleological explanations involve a distinctive loop between causes and effects. Walsh (2008) describes this teleological loop as follows: "teleology is a mode of explanation in which the presence, occurrence, or nature of some phenomenon is explained by the end to which it contributes" (Walsh 2008: 113). In a series of seminal papers in the 1970s, Wright (1976) proposed analyzing this explanatory loop by appealing to the causal history of the items to which purposes are attributed. In this view, the existence of an item can be accounted for in terms of some of its effects insofar as these effects play a crucial role in the causal history of the origin or preservation of the item. Wright thus inaugurates the so-called etiological approach to teleology, grounding purposes in causal history.

According to etiological approaches, the function of something is identified as the reason why it exists in its present form. Thus, in these approaches, to say that "the function of X is Y" is to say that "X exists because it does Y." Causal history (etiology) explains the presence of a feature through one of its effects, that is, its function. For example, the blood pumping function of the heart in the past explains why hearts exist today, or the function of the peacock's tail to attract mates in ancestral specimens explains the current existence of this trait. Functions have an explanatory role in accounting for the presence of these traits because of their historical relevance.

However, even if it is granted that the continued presence of a trait is explained by its tendency to produce certain effects, it is not immediately clear why this, on its own, would entail that the trait has a purpose and is subject to standards of success. There are multiple examples in the literature of cases where an entity has certain effects that de facto determine the continued existence of the entity without these effects actually being understood as "goals." For example, Bedau (1991) describes the case of a stick floating down a river that remains pinned on a rock due to the backwash it creates. As Bedau points out, "the stick does not create the backwash in order to keep itself pinned on the rock" (Bedau 1991: 648).

Etiological approaches can avoid this type of counterexample if they make use of selected-effects theories of teleology (Millikan 1984, 1989; Neander 1991; Griffiths 1993; Godfrey-Smith 1993). In these theories, the relevant teleological loop between causes and effects is generated by selective processes. That is, certain effects explain

the presence or proliferation of some traits because having these effects explains why those traits were selected in a relevant selective process. Accordingly, the purpose of a trait is doing whatever it was selected for.

In the example above, the position of the stick on the rock is not explained in terms of any selective process. Thus, selected-effects theories do not attribute purposes to the backwash of the stick. In general, it seems natural to employ teleological discourse whenever we find cases of selection. As Griffiths (1993: 420) claims, "where there is selection, there is teleology." The connection between selection and teleology is also highlighted by Neander (1991: 463): "Teleological explanations (. . .) explicitly refer to a future effect of a trait for which that trait was selected. In doing so they explain the trait by implicitly referring to the causally efficacious selection process from which it resulted."

Selection captures two of the central features of teleology: its evaluative nature and the existence of teleological loops. First, both teleology and selection have an evaluative dimension, understanding evaluative normativity as concerning what is good or bad (beneficial or detrimental) in some way (McLaughlin 2009). Purposes are associated with evaluative standards: a successful performance is good as an instance of purposeful behavior. Likewise, selection comes hand in hand with evaluation. We can think of selection as classification plus valence. Items selected *for* are positively evaluated with respect to the standards governing that selective process, while items selected *against* are evaluated negatively.

Moreover, selection gives rise to the type of teleological loop discussed in etiological theories like Wright's (1976). In selective processes, certain items are preserved, promoted, or positively reinforced in some way, whereas others are inhibited or negatively reinforced in some way. The reinforcement of selected items is explained by those effects that led to their selection.

Selected-effects theories offer, therefore, an attractive account of teleology. These theories deal well with many paradigmatic types of teleology. Consider intentional, goal-guided decision-making. This type of purposeful behavior involves selecting a course of action that contributes suitably to the achievement of the relevant goals, while discarding alternative actions that are detrimental to it. Because the relevant actions were selected due to their tendency to have certain effects or consequences, selected-effects theories regard these effects as the goals of such actions.

A second example is that of the functions of artifacts. The creators of artifacts select their features because of their role in producing certain effects. Therefore, according to selected-effects theories, the purposes or functions of artifacts are to produce these effects. For instance, the designers and manufacturers of hammers choose their shape (a compact head with a flat impacting surface) because of its usefulness to produce effects such as driving in nails. These are, therefore, the functions that creators of hammers intend them to have. Of course, if users select a hammer rather than another tool for a different task (say, being a paperweight), performing this further task can become the (perhaps temporary) function of the hammer in relation to the selecting intentions of those users. In the next section, we consider views that apply selected-effects theories to biology in terms of natural selection. We discuss the possibility of using natural selection as a model for a

general characterization of biological selection that covers other types of biological selective processes. We will argue that generalized accounts of (biological) selective processes modelled on natural selection are too restrictive and leave out legitimate kinds of biological selection, such as self-regulation.

3 Modelling Biological Selective Processes on Natural Selection

The vast majority of selected-effects theories in biology take natural selection as the basis of biological teleology (Millikan 1984, 1989; Neander 1991; Griffiths 1993; Godfrey-Smith 1993; Buller 1998, Artiga 2021). In this type of view, biological purposes are effects selected by natural selection. In this way, the existence of a current token of a purposeful trait is explained by the fact that past members of its lineage tended to produce certain effects, which led to their proliferation under natural selection.

An interesting question is whether selected-effects theories can be applied in biology beyond natural selection, that is, to other types of biological selective processes. One way to go here is to construct a general characterization of selective processes based on the features of natural selection (Darden and Cain 1989; Garson 2017). Natural selection would be taken as model for determining when a mechanism counts as selective and thereby grounds teleological explanations. The aim is then to find biological processes that fit this general characterization, by virtue of their being sufficiently analogous to natural selection. A reason to adopt this strategy is that natural selection is arguably the most studied and best understood biological selective process, so it makes sense to use it as a model to investigate further forms of biological selection.

Darden and Cain (1989) have proposed a generalized account of selection of this kind, inspired by the features of natural selection but applicable to other biological selective processes, such as clonal selection or immunological selection. In Darden and Cain's words:

> A selection process may be broken down into a series of steps from which a more abstract characterization can be developed.
>
> (A) First are the preconditions before a selective interaction. These include a set of individuals that vary among themselves. Also, the individuals must be in an environment with critical factors that provide a context for the ensuing interaction.
>
> (B) The actual step of selection involves an interaction between individuals and their environment. Because they vary, different individuals will interact differently.
>
> (C) Several types of effects result from the differential interactions. In the short-range, individuals benefit or suffer. If the individuals can be located in a hierarchy (such as gene, organism, group), then there may also be short-range effects of sorting at other levels.
>
> (D) Longer-range effects may follow the short-range effects of the interaction, such as increased reproduction of individuals with certain variations or reproduction of something associated with those individuals.

> (E) Even longer-range effects may also occur, such as accumulation of benefits through numerous generations to produce a lineage of individuals. (Darden and Cain 1989: 110)

Natural selection follows the steps of Darden and Cain's schema. In cases of natural selection, we have a population of individuals featuring variability in some of their properties (say, a population of peppered moths with different colors). Individuals in these populations interact with their environment in different ways depending on these variable properties (dark moths in trees darkened by soot are less visible to birds of prey). As a result of these different ways of interacting, individuals with some of the variable properties are more likely to survive and reproduce than others (dark moths are less likely to be eaten by said birds of prey). Thus, individuals with certain properties tend to proliferate. This leads to the formation of lineages in which some of the properties predominate (e.g., lineages of dark moths).

Darden and Cain argue that the core elements of evolutionary explanations can be generalized and applied in other selective explanations in biology. Explanations involving clonal or immunological selection, for instance, would be selective in the same way as explanations involving natural selection because they would share the same basic structure. As Darden and Cain (1989: 118–121) explain, in Burnet's (1957) clonal theory of antibody formation, we can observe the fundamental elements of selective processes: (A) There is a set of lymphocyte cells, with different reactive sites. (B) Different antigens activate lymphocyte cells with different reactive sites, making them produce clones of themselves. Thus, cells with different reactive sites interact in different ways with antigens in the environment. (C) In the presence of a given antigen, cells with a certain reactive site are activated and produce clones of themselves, while other cells are not activated. (D) Activated cells proliferate by cloning themselves and release antibodies of a specific type that attack the antigen that activated them. (E) More cells of this type are present after the antigen is eliminated, so that the immune system is able to respond more quickly to future invasions by that kind of antigen.

So, according to Darden and Cain's generalized characterization of selection, antibody formation involves selective processes which are different from natural selection, but with an analogous abstract structure. Darden and Cain's proposal clearly exemplifies what has been the most common strategy for investigating biological selective mechanisms. It is typical in philosophy of biology to assume that natural selection is the paradigmatic case of a selective mechanism in biology and, on the basis of this type of mechanism, to try to offer a generalization of the idea of selection that can be used to account for other biological processes. This strategy is followed, among others, by Griffiths (1993) and Garson (2017). We do not wish to deny that taking natural selection as a model can be useful in identifying and characterizing other types of selective processes in biology, such as immunological selection. However, we think that this approach leaves out interesting forms of biological selection that do not share some of the structural features of natural selection. The example we will focus on here is biological regulation. We argue that regulation deserves to be considered as a type of biological selection, at least as much as natural selection does, given that it shares the core features of paradigmatic

forms of selection. Accounts of selection modelled on natural selection are too narrow insofar as they exclude these paradigmatic types of non-biological selection. If we construct a characterization of selective processes broad enough to include these paradigmatic types of selection, biological regulation would also be included.

4 Selection Generalized

Garson's (2017) generalized account of selective processes mirrors the structure of natural selection less closely than Darden and Cain's. In particular, Garson argues that the forms of selection giving rise to functions can be a matter of differential retention, and not just of differential reproduction (as happens in natural selection).[1] Garson's main aim is to examine processes of neural selection, such as synaptic selection. Synaptic connections do not reproduce, but they can be differentially retained or eliminated in competitive processes, depending on their levels of activation. According to Garson, this form of competition can ground a form of selection, even in the absence of differential reproduction.

We consider Garson's view to be going in the right direction. There are many paradigmatic forms of selection without reproduction. Think of someone choosing apples in the supermarket. Obviously, the selected apples will not reproduce; they are just taken home by the buyer. Our proposal, however, is to go further than Garson and take selection to be more generally a matter of differential reinforcement. We understand reinforcement broadly, as including reproduction, retention, and different forms of promotion or enhancement. For instance, a way in which a process can be positively reinforced is by being stimulated or intensified. So, in glycemia regulation the release of insulin stimulates the absorption of glucose into muscle, adipose, and liver cells while suppressing the production of glucose through glycogenolysis. This is an example of differential (positive and negative) reinforcement of certain processes without reproduction or retention.

Our claim, therefore, is that selection involves the differential reinforcement of certain effects or traits, where this reinforcement may be a matter of being promoted, reproduced, preserved, stimulated, or intensified somehow (or, alternatively, inhibited, suppressed, or eliminated). The notion of reinforcement, we think, is sufficiently flexible to cover the great variety of cases of selection, including its most paradigmatic forms. It makes sense to consider natural selection a selective process precisely because it involves differential reinforcement, in the form of differential reproductive rates. Learning by trial and error is another example of selection based on reinforcement. In this case, the learner develops dispositions to repeat some behaviors and not others depending on whether they are observed to produce certain outcomes reliably.

[1] An account of selection in terms of retention can also be found in Campbell (1960).

There are other ways in which Garson's proposal remains, in our view, too attached to the model of natural selection. Garson (2017) requires that selection operates on a population of (actual) entities engaged in fitness-relevant interactions. This is certainly a feature of some forms of selection, but it is absent in many paradigmatic selective processes. Thus, a generalized account that includes this condition remains too restrictive and may leave out genuine cases of biological selection. Consider a process of selection of candidates for an academic distinction. Such a process may consist of an individual exam, without interactions among the candidates. Indeed, the process does not need to be competitive. Imagine that there is no limit to the number of distinctions that can be awarded. Then, whether a candidate gets the distinction is a matter of whether their exam results meet a given standard, regardless of how their performance compares to that of the other candidates.

It can be argued that selection does not always operate over a preexisting population of items featuring variability in their features. Go back to the academic distinction example. Think of a case where there is only one candidate. Still, it makes sense to say that, if the candidate passes the exam, they have been selected for the award. The important point is that the candidate has been selected for certain features or effects of their performance (in particular, their results in the exam) and would not have been selected if those effects had been different. Thus, selection has a modal dimension: selective processes involve dispositions to reinforce items with certain effects, so that the selection of some item is explained by its tendency to have those effects. This does not require there to be an actual population of items with different features. What matters is that should there be an item without the relevant effects, then it would not have been selected by the selective process in question. It is irrelevant whether such an item actually exists in any given population.

We can now sketch a rough generalized characterization of selective processes as driven by mechanisms with the disposition to differentially reinforce certain items by virtue of some of their effects or features. We are relying here on a basic notion of selective mechanism in line with most contemporary mechanistic approaches, such as that of Illari and Williamson (2012), for whom "a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon." It also fits what Glennan (2017) has called Minimal Mechanism: a mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized in such a way that they are responsible for the phenomenon. In the case of selection, the relevant phenomenon is the reinforcement of certain effects or traits over possible alternatives. Selective mechanisms are those entities and activities responsible for this type of phenomenon.

This is then our proposed general characterization of a selective mechanism:

> A selective mechanism is a mechanism by which the behavior of a system and its relationship with its environment are modified in such a way as to reinforce the presence of certain effects or traits over other alternatives.

We want to use this general abstract characterization to address in what sense certain explanations show a teleological dimension, making use of a selected-effects account of teleology. We will argue that an explanation is teleological if it appeals to effects of a trait that explain its reinforcement through a *selective process*.

We will therefore sustain this thesis about the selective character of teleology[2]:

> T1: A trait can be teleologically explained if it is structured as the result of a selective process.

We argue that a general characterization of the notion of selection provides a valuable abstraction of teleological explanation in all its diversity. This abstraction will help us not only in the task of naturalizing controversial teleological explanations but also in understanding how these explanations must be construed in order to be useful to current researchers in elaborating new theories.

In the next sections, we examine how our minimal, generalized account of selection applies to paradigmatic cases of intentional selection and also to natural selection. After that, we discuss biological regulation as an example of a biological process that is considered as selective according to our generalized account, despite not meeting some of the conditions of characterizations of selection modelled on natural selection. Our conclusion is that biological regulation should therefore be considered a selective process, giving rise to its own form of biological teleology.

5 Intentional Selection

Perhaps the most paradigmatic form of selection is that in which the relevant selective mechanism is constituted by an intentional agent who chooses among several possible options. This type of selection is found in explanations of typical human behavior, artifact design, and intentional selective breeding. Think, for example, of a customer selecting pieces of fruit in a supermarket, a committee selecting candidates for a position, or a family choosing a film to watch on TV. Similarly, the shape of a tool can often be explained by referring to the intentions of its designer. In this way, we can explain why a hammer has the shape it has (a compact head with a flat impacting surface) in terms of its intended use: the designer of the hammer intended it to perform tasks such as driving nails, so they selected a tool with a fitting shape for those purposes.

Methodical selective breeding (also known as artificial selection) is another paradigmatic example of intentional selection, studied in detail by Darwin (1868). In this type of intentional selection, a breeder chooses animals or plants with certain

[2]Note that this type of selected-effects theory posits sufficient, but not necessary, conditions of teleology. We leave open the possibility of teleology without selection.

phenotypical traits to reproduce together, promoting in this way the presence of those traits in their offspring.

In cases like these, an agent, or group of agents, makes intentional decisions by virtue of which certain options are selected. Given that these are paradigmatic instances of selection, they should be counted in by accounts that aspire to offer a generalized characterization of selection. If an analysis of selective processes imposes conditions that are not met in these paradigmatic instances of selection, then we have reason to think that such an analysis does not cover all central forms of selection. The fact that a mechanism does not satisfy the conditions of this analysis would not mean that it cannot count as selective.

Many cases of intentional selection fail to fulfil the conditions of characterizations of selective processes modelled on natural selection. In particular, intentional selection does not presuppose a preexisting population of items to be selected, nor fitness-relevant interactions, as Garson's (2017) account does. For instance, when choosing whether to go for a picnic or take a walk by the river, one does not select among preexisting picnics and walks, but rather chooses an option among possible (non-actual) alternatives. If I choose the walk, the picnic will remain an unrealized possibility. Moreover, intentional selection can involve a wide variety of forms of differential reinforcement, beyond differential reproduction or retention (we can also have, for instance, differential repetition or differential increases in the intensity or rate of a process).

But, of course, intentional selection does not stop counting as a selective process just because it does not fit in with generalizations of the notion of selection modelled after natural selection. Intentional selection is, if anything, a more paradigmatic type of selective process than natural selection. So, rather than denying the existence of intentional selection, authors like Darden and Cain tend to treat it as the basis of a different type of mechanistic explanation that deserves a separate analysis.

We grant that a pluralistic approach to selection mechanisms can be fruitful—after all, it is to be expected that these mechanisms will present a great degree of heterogeneity. However, in light of this pluralism, one should not assume that all relevant forms of biological selection share the distinctive structure of natural selection. Accounts based on natural selection are too narrow when taken as generalized characterizations of selective processes. That is why it can be useful to find some core characteristics of selection mechanisms that feature not just in natural selection but also in other central selective processes. Our proposal is that the notion of differential reinforcement allows us to develop just such a general characterization of selection mechanisms.

In particular, intentional selection can be perfectly captured by a characterization of selection in terms of differential reinforcement. In this case, an intentional agent (or a group of them) would be responsible for the differential reinforcement of the relevant items and would therefore act as a selective mechanism. We can now have intentional teleological explanations that are a variant of T1, where the selective mechanism is constituted by intentional agents:

T2: A trait can be teleologically explained if it is reinforced as the result of the selective process performed by intentional agents. [3]

6 Natural Selection

Intentional explanations have been very influential in the history of the life sciences. In particular, what we can call the intentionalist approach tries to account for biological teleology by appeal to the intentions of some agent. This type of approach is reflected in the famous design argument, wielded from deistic positions and vehemently defended for centuries by promoters of Natural Theology. In Natural Theology the Creator has devised the conformation and activity of living things. This powerful demiurge has "chosen," from a potentially infinite variety of alternatives, the design of the concrete organization of each biological individual with all its particularities. Thus, natural organizations are teleological because they respond to a specific purpose, which would be the purpose for which their creator has designed them in such a way. [4]

In naturalistic approaches to biological teleology, the action of a supranatural intentional selector is left out of the explanation, but the primary role of selection in grounding biological teleology has been, to a large extent, retained. In etiological-evolutionary views, predominant in the current philosophy of biology, it is the action of natural selection that confers purposes to biological traits. Thus, biological purposes would no longer be the impositions of an external intentional selector, but the result of a long evolutionary history in which certain effects have been preserved and others have disappeared. Natural selection "chooses" biological purposes.

Consequently, from a completely different starting point than intentionalist approaches, etiological-evolutionary theories also consider selection to be the basis of teleology. In this case, the relevant selective mechanism is natural selection (see Barros 2008 for a defense of the view of natural selection as a [stochastic] mechanism). Differential reinforcement here takes the form of differential reproduction, so that selected traits are those that proliferate under selective pressures. It is thus

[3] This formulation of teleological intentionalism has clear precedents in Broad's classic proposal (Broad 1925: 82).

[4] For a detailed analysis of the "design argument," see Sober (2018). Put forward by Hume in his *Dialogues Concerning Natural Religion*, and famously expounded in the the early nineteenth century by William Paley, this argument has been pervasive in the history of biology up to the present day.

possible to speak of a "what for" [5] in biology because we can identify a selective mechanism, natural selection, that explains the actual existence (more specifically the proliferation) of certain traits and their effects. In this way, etiological-evolutionary approaches can be seen as a particular case of T1:

> T3: A (biological) trait can be teleologically explained if [6] its proliferation is the result of natural selection.

Natural selection is the key to grounding teleology in a naturalistic approach because it is understood to be a selection mechanism, in the same way that intentional, mentalistic selection mechanisms can be identified in cases of intentional purposive behavior. Nature selects an effect of a trait in the analogous way that we intentional agents choose from among different alternatives (we choose courses of action, friendships, or supermarket products). To be sure, these two types of selection mechanisms are very different in their specific details, but both exhibit the core, characteristic feature of selective processes: differential reinforcement. Take, for example, the standard etiological explanation that the function of the heart is to pump blood because this is the effect responsible for the preservation of this trait throughout the evolutionary history of mammals under the pressure of natural selection (Buller 1999: 1–7). This explanation treats natural selection as the selection mechanism underlying ascriptions of functions. In our proposal, natural selection can be considered as a selective mechanism insofar as it involves a historical form of differential reinforcement. So, hearts that fail to pump blood tend not to proliferate; their presence in future generations tends to be inhibited.

One could try to argue that, when investigating biological teleology, we can focus on those forms of selection typical in biology. What kinds of selection are paradigmatic in non-biological domains would not be relevant, so there would be no need to find a generalized account that covers non-biological selective processes. So, we could have a splitting account that distinguishes different types of selective mechanisms, giving rise to different types of teleology, without a general account of selection unifying them.

This splitting account, however, is problematic. First, it may undermine the application of selected-effects theories of teleology in the biological domain. The connection between selection and teleology is especially clear in paradigmatic cases of non-biological selection, in particular in intentional selection. If one holds that biological selection does not belong to a common kind with other types of selection, then the idea that biological selection introduces teleology may be questioned. One

[5] A classic discussion of the distinction between finalistic "what for?" questions, as opposed to historical "how come?" ones, can be found in Mayr (e.g., 1961).

[6] Again, this is only a sufficient but not a necessary condition. A trait can be explained teleologically in relation to its selective history, but not only so, as shown, for example, by teleological explanations based on biological regulation.

objection could be that we are equivocating different notions of selection: we start with a connection between teleology and non-biological selection, and then we export it illegitimately to the biological domain, despite the fact that biological selection and non-biological selection are different types of phenomena. It could even be claimed that the term "selection" is used in biology only in an extended or metaphorical sense, which should not carry teleological implications that are only warranted in genuine cases of selection. All these worries will be assuaged if we can show that biological selection shares the central features of those forms of selection in which the connection with teleology is undeniable. So, a generalized characterization of selection covering biological and non-biological cases puts biological selected-effects theories on a firmer footing.

Moreover, as we have already pointed out, there are interesting biological selective processes that do not fit the mold of natural selection, such as biological regulation. So, even if we are only interested in biological teleology, a characterization of selection with natural selection as its model will remain too restrictive. We need a more general account that also captures these other types of biological selection.

In the next section, we discuss biological regulation as an example of a biological selective process that satisfies our generalized characterization of selection, despite not mirroring the structure of natural selection. We will argue that, in many respects in which the features of regulation and of natural selection diverge, the former is closer than the latter to paradigmatic forms of selection. So, discussing selection in relation to regulation is not more of a stretch (if anything, less) than doing so in relation to natural selection.

7 Regulation

Besides the evolutionary approach, there is another etiological tradition in philosophy of biology: the so-called cybernetic approach, based on ideas introduced by Rosenblueth et al. (1943) and developed, among others, by Sommerhoff (1950, 1959). The cybernetic theorists locate a teleological loop *á la* Wright in concrete dynamics that occur in the framework of the current organization of living beings. In particular, authors in this current have tended to consider that the feedback mechanisms that guarantee the stability of biological systems, such as homeostasis, provide the basis for the attribution of biological functions and, consequently, for teleological discourse in biology (Adams 1979; Boorse 1976; Edin 2008). Cybernetics is also one of the pillars on which the current organizational approach to biological teleology is based (Christensen and Bickhard 2002; Mossio et al. 2009; Saborido et al. 2011; Schlosser 1998). According to this approach, a trait function would be the effect of this trait that makes a contribution to the dynamic maintenance of the conditions, both internal and relational, which allows the living system to continue existing. For example, the function of the heart would be the pumping of blood because this pumping of blood has effects that have a direct impact on biological

self-maintenance, such as the transport of nutrients to the cells, the stabilization of temperature and pH, and so on. The cybernetic approach allows for a naturalized explanation of certain teleological statements of biological systems based on the properties of their organization.

In recent years, the notion of biological regulation has become particularly relevant for the study of the organization of living beings, offering a more sophisticated approach to biological self-maintenance processes than other notions traditionally used by cybernetic approaches. In line with Bich et al. (2016), [7] regulation can be characterized as the capability to actively modulate the internal dynamics and behavior of a system in relation to variations in internal and external conditions. Regulation is the result of specialized mechanisms that evaluate disturbances and operate accordingly. These regulatory subsystems are "sufficiently independent of the dynamics of the controlled processes, and which can be varied without disrupting these processes, but it is still able to be linked to parts of the mechanism controlled system [the regulated subsystem] so as to be able to modulate their operations" (Bechtel 2007: 290).

Regulatory mechanisms are central to biological organization and are discussed in detail in the biological sciences, as the example of lac operon shows. Thus, as described by Bich et al. (2016), in the case of lac operon, two subsystems are identified: "the regulatory subsystem (consisting of the DNA sequence -promoter, operator, genes- plus regulatory proteins) and the regulated one (metabolism, or parts of it)" (Bich et al. 2016: 261). The lac operon is a concrete mechanism of regulation of protein synthesis, i.e., a process by which a cellular regulatory subsystem of an organism is able to choose what proteins or enzymes to produce given certain environmental characteristics, such as the availability of specific amino acids. Therefore, a system with adaptive regulation is capable of actively modulating its internal dynamics and behavior. It is not simply a matter of resilience or robustness, in which an organism passively "resists" the pressure of the environment. Regulation, instead, enables the organism to actively engage with the environment through selective processes: regulation involves a mechanism of selection—the regulatory mechanism—of the appropriate operations that a biological system must perform given specific circumstances.

Some simple types of regulation take the form of homeostatic stability. Homeostasis refers to the capacity of certain systems to maintain their internal dynamics in a stable attractor in the face of external perturbations or internal variations. Regulatory mechanisms can sustain homeostasis by adjusting the behavior of the system to these perturbations and variations, so that the stability of the system is preserved.

For example, mammals are able to maintain blood sugar levels within fairly narrow limits. If there is a rise in blood sugar, the beta cells of the pancreatic islets

[7] In this paper we focus mainly on the characterization of regulation developed by Bich and collaborators and presented in Bich et al. (2016), Bich et al. (2020), and Bechtel and Bich (2021), because it is a particularly well-developed analysis of biological regulation, particularly well suited for conceptual and philosophical analysis (see also Winning and Bechtel 2018).

respond by secreting insulin into the blood and, at the same time, preventing their neighboring alpha cells from secreting glucagon into the blood. The combination of a high level of insulin in the blood and a low level of glucagon triggers the action of the effector tissues, mainly the liver, fat cells, and muscle cells, which—through different physiological mechanisms of inhibition and glucose uptake—manage to correct the excess glucose in the blood. On the other hand, if the perturbation faced by the mammal is a drop in blood glucose, this causes the interruption of insulin secretion and glucagon secretion from the alpha cells into the blood. Here, the uptake of glucose from the blood by the liver, fat cells, and muscles is inhibited, and, instead, the liver is strongly stimulated to produce glucose, which is discharged into the blood, thus reversing the hypoglycemia.

In the case of glycemia regulation, the relevant regulatory subsystem would involve beta and alpha cells in pancreatic islets, which detect significant variations in blood sugar levels and carry out actions that have an impact on other parts, such as the liver, fat cells, and muscles, ultimately leading to homeostasis, that is, the return to the original state in which the blood sugar level is within certain variables (Bich et al. 2020: 9).

This example shows that homeostasis can be an instance of regulation, insofar as it is achieved through the intervention of a regulatory mechanism. However, it is important to bear in mind that not all forms of regulation are homeostatic. Approaches that claim that homeostasis is a goal of the biological systems presuppose that the viability of the system depends primarily on its stability (Ashby 1956; Keller 2008). In this view, biological organisms would be systems that need to be brought back to their default state when they are affected by perturbations, as shown by the case of blood sugar levels. By contrast, regulatory adaptability does not necessarily imply a return to this default state. Adaptability implies change. Although the activity of regulatory mechanisms may result in the stability of some variables, this homeostasis cannot be considered as a goal in itself, but as a means to maintain the viability of the system. Indeed, in several cases, the viability of the system is achieved precisely by moving away from the original state. Considering that living systems are continuously interacting in a changing environment and undergoing internal transformations, regular behavior and stability might be the exception rather than the rule. Different organisms, or the same organism in different moments, may exhibit different set points for their physiological variables.

Let us consider a simple example. In a situation of danger, heartbeat and blood pressure change with respect to a situation of rest to provide oxygen and glucose for skeletal muscles. Regularity and stability in this case would not be beneficial but rather detrimental to the self-maintenance of the organism. An adaptive response to the situation of danger requires changes in the behavior of the heart, for instance, a faster heartbeat (Saborido and Moreno 2015).

Thus, instability can be more adaptive than stability, as the extremely large number of degrees of freedom it provides makes it possible for the system to carry out appropriate operations in extremely varied situations (Kitano 2007). Regulatory adaptivity focuses on engaging with, and taking advantage of, variability and change, instead of preventing them. Regulatory mechanisms do not only respond

to perturbations that threaten the survival of the system or destabilize some variables in the system. A system endowed with adaptive regulatory mechanisms can implement new types of organization based on the environmental conditions. This is already observable in very simple cases of biological organization, for example, in bacteria, to actively exploit opportunities in the environment rather than merely react to it: to follow a gradient of concentration of nutrients, to synthesize different enzymes to metabolize different nutrients depending on their availability in the environment and their energy efficiency, to establish themselves in given locations or to move to others, etc.

The regulatory modulation of the dynamics of organisms constitutes a form of differential reinforcement. In general, regulatory mechanisms promote certain states or dynamic tendencies of the organism while inhibiting others. In some cases, this may be a matter of preserving the homeostatic stability of the organism, actively suppressing or counteracting deviations from such stability. Deviations, therefore, tend to be counteracted by regulatory mechanisms (for instance, in order to keep a certain level of blood sugar). But, as we have seen, regulatory mechanisms sometimes switch between different modes of operation of the organism, as when a heartbeat gets faster in response to danger. Here the regulatory mechanism triggers the realization of one possible mode of behaviors of the regulated system over the alternatives (in this case, different heartbeat frequencies).

In accordance with our characterization of selection, regulation can be regarded as a selective process, by virtue of the role that differential reinforcement plays in it. Regulatory subsystems act as selective mechanisms that are disposed to differentially reinforce certain states and tendencies of the organism. In a selected-effects approach, therefore, biological regulation grounds teleological explanations based on the organizational properties of living beings.

Note that this is so even if regulatory selection does not satisfy the conditions of accounts of selection developed by generalizing the structure of natural selection, such as Darden and Cain's (1989). Among other things, regulation does not typically operate over a preexisting population of different items with variable features (for instance, a population of different heartbeats). Despite this, it makes sense to count regulation as a form of selection mechanism, since it exhibits the core features of selection, in particular differential reinforcement leading to a loop between effects and causes.

It is worth mentioning one aspect in which regulation is closer than natural selection to central instances of selection. In most paradigmatic cases of selection, it is possible to identify a selector, a concrete agent or system responsible for the reinforcing pressures driving the selective process. One can point to such selectors in order to answer the question about who performs the relevant selection. For example, in a recruitment process, the selector is the hiring committee. Likewise, regulatory subsystems play the role of selector in biological regulation. Remember that regulatory mechanisms are constituted by concrete, physically realized systems in the organism. By contrast, in natural selection it is not clearly the case that one can find a concrete, individuated selector. To the question of who selects in natural selection, one can only offer vague answers—perhaps nature—but it is hard to find a

concrete entity responsible for such selection. We do not intend this to discredit natural selection as a selective process. We grant that there is a well-defined notion of mechanism according to which natural selection can count as a (selective) stochastic mechanism (see Barros 2008). However, this example shows that biological regulation is in some ways more similar to paradigmatic selective processes than natural selection.

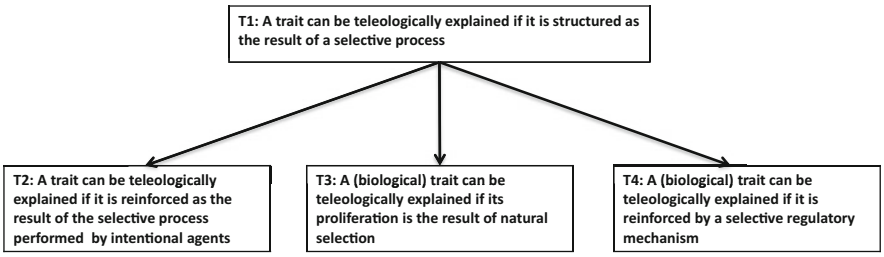
Our claim, therefore, is that it is possible, on the basis of the notion of biological regulation, to develop a naturalistic, etiological account of biological teleology that is different from (even if compatible with) evolutionary approaches. Regulation grounds teleological explanations because it is a selective process, by which biological systems modulate the conditions of their interaction with the environment in order to continue to maintain themselves. Regulatory selective mechanisms allow us, therefore, to naturalize certain teleological explanations that appeal to the functioning of biological organizations "here and now" (appealing to what Mayr (1961) defined as "proximate causes"). In this way, we get a further application of our selected-effects schema T1, this time associated with regulatory selection mechanisms:

> T4: A (biological) trait can be teleologically explained if it is reinforced by a selective regulatory mechanism.

8 Conclusions

In the selected-effects approach we have pursued here, teleological explanations are justified by their appeal to minimal selective mechanisms, that is, mechanisms that involve differential reinforcement. The existence of a mechanism that selects certain effects gives rise to teleological loops, in which the presence of a trait is explained by some of its effects. Selection, consequently, is a source of teleology in all domains in which it takes place.

In this way, teleological explanations based on the intentions of rational agents (T2), on the action of natural selection (T3), and on the regulatory mechanisms of biological organizations (T4) are particular cases of selective teleological explanations (T1), focusing on different selective mechanisms.



T2 has no place in naturalistic explanations in biology—at least, unless we give a naturalistic account of the relevant intentional selectors. On the other hand, biological traits can often be explained from the perspective of both T3 and T4, given that the same trait can be under the scope of both natural selection and regulation. These explanations are compatible in a strict sense, since they just focus on different selective regimes to which the item is subject to. In many cases these different selective processes will promote the same effects, so that the function ascribed to a trait will be the same from both approaches. For example, pumping blood is the function of the heart both because it is an evolutionarily selected effect and because it is also the result of organismic regulation. In other cases, this is less clear. For example, functions can be identified and teleological explanations offered for traits that have not yet undergone the action of natural selection (emergence of new functions) or that have changed their function at some point in their evolutionary history (exaptations).

The selective actions of regulation and of natural selection are interdependent: natural selection acts on traits whose behavior is modulated by the regulatory mechanisms of individual organisms, and, at the same time, these regulatory mechanisms and the traits they regulate have been shaped by natural selection.

Teleological explanations based on regulatory mechanisms are therefore not intended to replace teleological explanations based on the mechanism of natural selection. Both share the same logical structure: the action of a selection mechanism justifies the ascription of purposes. If teleological explanations have any scientific value, it should be their contribution to increasing knowledge about biological phenomena. A naturalistic approach to biological teleology will therefore benefit from the inclusion of explanations that take into account selective mechanisms other than natural selection which, like regulation, serve to provide a better insight into the reasons why certain biological structures and processes originate and are preserved.

References

Adams FR (1979) A goal–state theory of function attributions. Can J Philos 9(3):493–518
Artiga M (2021) Biological functions and natural selection: a reappraisal. Eur J Philos Sci 11(2): 1–22
Ashby WR (1956) An introduction to cybernetics. Chapman and Hall, London

Barros DB (2008) Natural selection as a mechanism. Philos Sci 75(3):306–322
Bechtel W (2007) Biological mechanisms: organized to maintain autonomy. In: Boogerd F, Bruggerman F, Hofmeyr JH, Westerhoff HV (eds) Systems biology. Philosophical Foundations, Elsevier, Amsterdam, pp 269–302
Bechtel W, Bich L (2021) Grounding cognition: heterarchical control mechanisms in biology. Philos Trans R Soc B: Biol Sci 376(1820):20190751
Bedau M (1991) Can biological teleology be naturalized? J Philos 88(11):647–655
Bich L, Mossio M, Ruiz-Mirazo K, Moreno A (2016) Biological regulation: controlling the system from within. Biol Philos 31(2):237–265
Bich L, Mossio M, Soto AM (2020) Glycemia regulation: from feedback loops to organizational closure. Front Physiol 11
Boorse C (1976) Wright on functions. Philos Rev 85(1):70–86
Broad CS (1925) The mind and its place in nature. Routledge and Kegan Paul, London
Buller DJ (1998) Etiological theories of function: a geographical survey. Biol Philos 13(4):505–527
Buller DJ (ed) (1999) Function, selection, and design. SUNY Press, Albany NY
Burnet FM (1957) A modification of Jerne's theory of antibody production using the concept of clonal selection. Aust J Sci 20:67–69
Campbell DT (1960) Blind variation and selective retentions in creative thought as in other knowledge processes. Psychol Rev 67(6):380
Christensen WD, Bickhard MH (2002) The process dynamics of normative function. Monist 85(1): 3–28
Darden L, Cain JA (1989) Selection type theories. Philos Sci 56:106–129
Darwin C (1868) The variation of animals and plants under domestication, vol 1 and 2. John Murray, London
Edin B (2008) Assigning biological functions: making sense of causal chains. Synthese 161:203–218
Garson J (2017) A generalized selected effects theory of function. Philos Sci 84(3):523–543
Glennan S (2017) The new mechanical philosophy. Oxford University Press, Oxford
Godfrey-Smith P (1993) Functions: consensus without unity. Pac Philos Q 74:196–208
Griffiths PE (1993) Functional analysis and proper functions. Br J Philos Sci 44(3):409–422
Illari PM, Williamson J (2012) What is a mechanism? Thinking about mechanisms across the sciences. Eur J Philos Sci 2:119–135
Keller EF (2008) Organisms, machines, and thunderstorms: a history of self-organization, part one. Hist Stud Nat Sci 38(1):45–75
Kitano H (2007) Towards a theory of biological robustness. Mol Syst Biol 3(137):1–7
Kitcher P (1993) Function and design. Midwest Stud Philos 18(1):379–397
Mayr E (1961) Cause and effect in biology. Science 134:1501–1506
McLaughlin P (2009) Functions and norms. In: Krohs U, Kroes P (eds) Functions in biological and artificial worlds. Comparative philosophical perspectives. MIT Press, Cambridge, pp 93–102
Millikan RG (1984) Language, thought, and other biological categories: new foundations for realism. MIT Press
Millikan RG (1989) In defense of proper functions. Philos Sci 56(2):288–302
Mossio M, Saborido C, Moreno A (2009) An organizational account for biological functions. Br J Philos Sci 60(4):813–841
Neander K (1991) Functions as selected effects: the conceptual analyst's defense. Philos Sci 58(2): 168–184
Rosenblueth A, Wiener N, Bigelow J (1943) Behavior, purpose and teleology. Philos Sci 10:18–24
Saborido C, Moreno A (2015) Biological pathology from an organizational perspective. Theor Med Bioeth 36(1):83–95
Saborido C, Mossio M, Moreno A (2011) Biological organization and cross-generation functions. Br J Philos Sci 62(3):583–606
Schlosser G (1998) Self-re-production and functionality: a systems-theoretical approach to teleological explanation. Synthese 116:303–354

Sober E (2018) The design argument. Cambridge University Press, Cambridge
Sommerhoff G (1950) Analytical biology. Oxford University Press, Oxford
Sommerhoff G (1959) The abstract characteristics of living organisms. In: Emery FE (ed) Systems thinking. Harmondsworth, London, pp 147–202
Walsh DM (2008) Teleology. In: Ruse M (ed) The Oxford handbook of philosophy of biology. Oxford University Press, Oxford, pp 113–137
Winning J, Bechtel W (2018) Rethinking causality in biological and neural mechanisms: constraints and control. Mind Mach 28(2):287–310
Wright L (1976) Teleological explanations: an etiological analysis of goals and functions. University of California Press, Berkeley

Evolutionary Causation and Teleosemantics

Tiago Rama

Abstract Disputes about the causal structure of natural selection have implications for teleosemantics. Etiological, mainstream teleosemantics is based on a causalist view of natural selection. The core of its solution to Brentano's Problem lies in the solution to Kant's Puzzle provided by the Modern Synthesis concerning populational causation. In this paper, I suggest that if we adopt an alternative, statisticalist view on natural selection, the door is open for two reflections. First, it allows for setting different challenges to etiological teleosemantics that arise if a statisticalist reading of natural selection is right. Second, by providing a different solution to Kant's Puzzle based on individual causes of evolution, statisticalism promotes a different answer to Brentano's Problem, what I label as Agential Teleosemantics.

Keywords Evolutionary causation · Causalist vs. statisticalist · Etiological teleosemantics · Agential Teleosemantics · Biological agency

1 Introduction

In the first paragraph of most papers about teleosemantics, this approach is usually introduced as the main attempt to naturalize intentionality. The reason is that its proposal is rooted in a biological account of natural teleology. This work is about such roots. I approach this issue from current debates between the statisticalist and causalist views on the causal structure of natural selection. I aim to argue that the statisticalism vs. causalism debate has important consequences for teleosemantics. Here I adopt a statisticalist viewpoint. The implications are twofold: a critique of teleosemantics and a reconstruction of it. First, it signifies different challenges to etiological (mainstream) teleosemantics—insofar as it is based on a causalist view of natural selection. Second, the statisticalist view allows for the development of an

T. Rama (✉)
Autonomous University of Barcelona, Barcelona, Spain
e-mail: tiago.rama@e-campus.uab.cat

J. M. Viejo, M. Sanjuán (eds.), *Life and Mind*, Interdisciplinary Evolution Research 8, https://doi.org/10.1007/978-3-031-30304-3_14

alternative solution to the naturalization of intentionality based on a different view on the causes of adaptive evolution.

A central issue around the statisticalism vs. causalism debate concerns the explanatory role of individual organisms in evolution. As remarked by many, the Modern Synthesis has managed to eliminate organisms from their explanation of evolution. This gave rise to black-boxing development (Hamburger 1980). Accordingly, evolution can be explained by looking into sub-organismal units of replication (genes) being arranged in supra-organismal units (populations) adaptively biased by natural selection. Organisms are not part of the picture. Even though the statisticalism vs. causalism debate concerns the nature of populational explanations, whether they are causal or statistical, different positions regarding the explanatory role of organisms in evolution provide support for either side.

I outline two implications for teleosemantics from a statisticalist viewpoint: challenging teleosemantics and resetting it. I propose two readings of these implications. The *weak implication* is formulated as a conditional: *if* the statisticalist view is correct, *then* teleosemantics must be challenged and reframed from a statisticalist perspective. The *strong implication* is to argue for the antecedent of the conditions, i.e., that the statisticalist view is correct, and, therefore, to defend the consequent. I will support the strong implication, even though I find the weak implication a valuable analysis to situate teleosemantics in debates about the nature of natural selection. As I will argue towards the end of the paper, support for statisticalism comes from recent proposals in organismal agency and its roots in individual-level causation.

In Sect. 2, I start by introducing Brentano's Problem on intentionality and Kant's Puzzle on teleology, in order to present the core of the teleosemantic project and the form it has been taking since its origin. In Sect. 3, I present the causalist view and its relation with etiological teleosemantics. In Sect. 4, I move to a different thesis concerning the causes of natural selection, what is known as the Statisticalist School. After presenting it in detail, I put forward two challenges to etiological teleosemantics that arise if we accept statisticalism. Finally, in Sect. 6, I present an alternative teleosemantic project motivated by the Statisticalist School, where the causes of evolution are seen as ontogenetic in nature and related to the *adaptive agency* of individual organisms. I conclude with an outline of the core tenets of *Agential Teleosemantics.*

2 Brentano's Problem through Kant's Puzzle: Setting Teleosemantics' Core

2.1 Brentano's Problem and Kant's Puzzle

Brentano's Problem is a good starting point to understanding the contemporary issues about the naturalization of intentionality. Intentionality concerns the capacity

of certain natural states (paradigmatically, cognitive states) to be about or refer to something else. Brentano's view placed intentionality in a paradoxical situation. Intentional explanations are explanatory useful and indispensable to understand goal-directed behavior, yet intentional explanations seem not to be aligned with the foundations of modern science. This implied, as Brentano emphasized in his *Psychology from an Empirical Standpoint*, that intentionality cannot be naturalized. There cannot be a science of goal-directed behavior that involves intentionality.

The issue turns around causation. To see this clearly, let's present the contemporary version of Brentano's Problem under the view of cognitive science and the philosophy of mind. In a nutshell, the mainstream view since the cognitive revolution in the mid-twentieth century posits that goal-directed behavior could be explained by appealing to the processing of representations. The idea is that animals are able to represent the world in a certain way. Such representations are achieved by complex mechanisms of perception and categorization. The information reached about the environment is processed in such a way that the animal behaves according to both its representation of its environmental circumstances and its inner goals (desires, needs, emotions, etc.). Since then, different disciplines within cognitive science have emerged incorporating some form of representationalist talk, and their inquiries are devoted to understanding how animals are able to represent the world, process such contentful information, and behave intelligently and adaptively.

Note first, as Brentano did, that intentional states could be about, for instance, nonexistent objects or future scenarios. However, if we explain behavior based on how the organism responds to the world, it seems difficult to figure out how we can provide causal explanations involving nonexistent objects. Shortly, the trouble arises when the causal chains underpinning certain goal-directed behavior involve a nonexistent object. Nonexistent objects are not part of causal chains; nonexistent things have no causal powers, but it seems that representations of such nonexistent objects do have causal powers. This is usually presented as a mismatch between the existence of the intentional object and the existence of the world object; that is, there could be an intentional object (a representation) without reference. This mismatch needs to be accounted for insofar as it seems that representations of nonexistent entities are necessary to explain certain behaviors. The principal trouble connected with this mismatch is latent in the issue of misrepresentation—a highly discussed issue in the teleosemantic literature. Intentional states could misrepresent the world. We have false beliefs, hallucinations, perceptual errors, conjectures, imaginations, false scientific theories, and so on. Even in the simplest cases, such as the famous case of frogs trying to catch flies, misrepresentation is present. In this sense, we can say that intentionality transcends what is actual. Let's call Causal Mismatch the mismatch between the intentional object and the world object, insofar as such relationships seem not to be suitable to be understood in terms of causal chains.[1]

[1] It is relevant to remark that the connection between causation and misrepresentations takes place within teleosemantics' aim of naturalizing intentionality. In this sense, teleosemantics was born against two classical approaches to representational content. One is the idea that the content of a

Brentano's Problem (under a contemporary reading) arises when we appreciate the following two points: First, intentionality explains. Intentional states are systematically used to predict, understand, and systematize goal-directed behavior in a truthful way. These kinds of explanations are usually labeled as *Folk Psychology*. However, to refer to them, I will opt for the name *Folk Intentionality,* to emphasize the parallelism with *Folk Teleology* (cf. below) and to provide a broader view of Brentano's Problem without assuming that whatever is intentional is also psychological. Moreover, since the beginning of the cognitive revolution in the mid-twentieth century, practically the whole cognitive enterprise is grounded in *Folk Intentionality* in a way that intentional talk plays a central explanatory role in any of the subfields of cognitive science—this includes also nonclassical accounts of intentionality within the so-called post-cognitivism, but clearly not some eliminativist positions concerning intentionality. However, secondly, the Causal Mismatch is not present in other sciences like physics or chemistry. These sciences were taken as the paradigm of scientific progress and the foundational roots of modern sciences since Descartes. Supposedly, in physics and chemistry, there are no gaps in the causal connections that lead to the *explanandum*. Step-by-step causal interactions, even if complex, upward or downward, produce the phenomenon to be explained. Here lies what is sometimes labeled as the Explanatory or Causal Asymmetry principle in science (Bromberger 1966; Potochnik 2017): step-by-step causal chains go from the past to the present, from the *explananda* to the *explanandum*. We can, therefore, explain the behavior purely based on neurophysiological causation, i.e., on those causal processes taking place in the nervous systems that step by step produce a particular behavioral outcome. However, such kinds of neurophysiological explanations would not involve any sort of intentionality, and consequently the behavior explained could not be treated as goal-directed.

Brentano's Problem lives in this contradictory scenario. On the one hand, intentionality is explanatory central in the cognitive enterprise, but intentional states cannot be understood under the same foundations of modern science. In other words, if causal interactions between nerve cells produce a certain behavioral outcome, and (ontological dualism aside) intentional states are neural states, how is it possible that intentional explanations give rise to the Causal Mismatch?

---

representation is determined by its relation with other representations – known as intensionalist theories. The difficulty of this attempt is naturalism, insofar as we cannot be naturalists and simultaneously anchor intentionality in other intentional stuff. In this view, misrepresentation cannot be a problem of causation insofar as content is not determined by reference but by how representations relate with each other. Teleosemantics opposes this view; it provides a referentialist theory of content. However, the second theory that teleosemantics rejects is the causalist one, which is also referentialist. Even if causalist theories of content pursue a naturalistic view of intentionality, they seem incapable of dealing with misrepresentations. In this sense, teleosemantics emerged as the attempt to solve misrepresentation and simultaneously accept the causalist view of intentional explanations defended by causalist theories. Therefore, misrepresentation becomes primarily an issue of causation once we stress the commitment of teleosemantics to naturalism. I thank an anonymous review for noting the importance of this clarification.

Kant's Puzzle concerns teleology. Its structure parallels that of Brentano's Problem. I am not going to expand very much here because I will retake this issue in the next section. Teleology concerns the capacity of certain systems to have a certain goal or purpose to fulfill. Teleological explanations in biology, therefore, appeal to what a certain system pursues in order to account for the functioning, activities, and behavior of such a system. Kant's position concerning teleology parallels Brentano's position concerning intentionality. For Kant, organisms cannot be understood without appealing to teleology. We cannot understand how organisms work, their complex organization, and their aptness to the environment eschewing teleological talk. However, he did not believe that teleological explanations could be a genuine part of science. According to Kant, in his *Critique of Judgment*, there cannot be science for purposive natural systems. Why is this so?

The reason, again, turns around causality. The kinds of explanations involved in teleology seem not to be aligned with the kinds of explanations accepted in science, even though teleology appears to be inexorable in our understanding of nature. To appreciate the paradoxical situation, let's remark first that teleology, just like intentionality, offers us, and scientific practice, a useful and veridical way to predict, explain, systematize, and interact with living beings. To unify terminology, I will call *Folk Teleology* these kinds of explanations. The second point to recognize is the status of the causal bases posited in teleological explanations. As noted, teleological explanations appeal to purposes and goals. Nonetheless, such purposes concern future stages, possible outcomes to be obtained. How is it possible that future stages explain current activities? Here causation enters the scene. An outcome cannot cause those processes that produce it. While, as explained before, science is built from step-by-step forward causal interactions—in biology, the psychochemical causes of physiological processes—teleological explanations seem to involve backward causation. This character of teleological explanation violates the Explanatory or Causal Asymmetry principle. The causal interaction underpinning any scientific explanation must run in one direction, from previous states of affairs to future states of affairs, not the other way around; the asymmetry regards the fact that causal interactions—even if complex, upward or downward—are unidirectional in time.

We can therefore see Kant's Puzzle also as a problem of naturalization. As with the case of intentionality, there is a tension between what is explanatorily desirable (Folk Teleology) and an uncomfortable situation with the causal basis of scientific explanations (Causal Asymmetry). The project of naturalizing teleology aims at figuring out how teleology can be legitimated by science without involving old-fashioned causation.

We can appreciate many connections between intentionality and teleology. Particularly, here I shall focus on problematic aspects of both phenomena. These connections are presented in Fig. 1. First, both riddles—the problem and the puzzle—take place when we appreciate that intentionality and teleology are explanatory necessary but still explanatory illegitimate. This creates a gap between the kinds of explanations and causations reputable in science (Causal Mismatch and Causal Asymmetry) and the advantages of assuming teleological and intentional

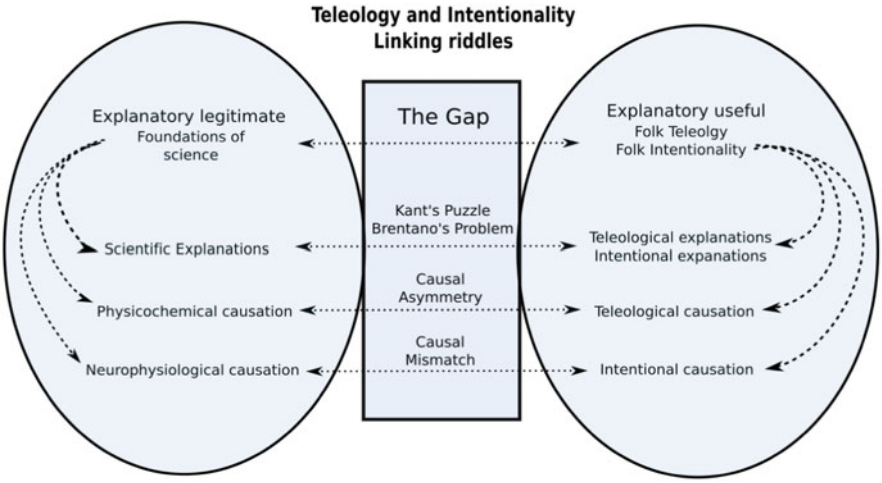


Fig. 1 The connections between Kant's Puzzle and Brentano's Problem. Both concern a tension between useful and legitimate explanations according to the foundations of science. This gives rise to a gap that naturalist projects aim to bridge. It consists in explaining how teleological causation could be compatible with physicochemical causation such that the principle of Causal Asymmetry stands and accounting for intentional causation in terms of neurophysiological causation to solve the Causal Mismatch

positions (Folk Intentionality and Folk Teleology). Is it possible to bridge this gap in naturalistic terms without throwing the baby out with the bathwater?

2.2 Teleosemantics' Core and Etiological Teleosemantics

Teleosemantics' core is to provide a naturalist solution to Brentano's Problem on the basis of *any* naturalist solution to Kant's Puzzle.[2] This involves two important theses. First, intentional states have teleological functions: any cognitive phenomenon performs a certain task directed towards the fulfillment of a certain goal. Second, intentional states, as natural phenomena, can be analyzed with the same tools as other natural traits, in a way that the analysis of teleological functions done in biology also encompasses that of cognitive functions. These theses are not going to be discussed here insofar as they are generally accepted within teleosemantics.

[2]Certainly, there could be a non-naturalist teleosemantic project. For instance, if we adopt a theological view of organism design, we can define the function of representational systems but in this case not from a scientific perspective. Moreover, as pointed out by an anonymous reviewer, there might be an attempt to solve Brentano's Problem from the notion of biological function with any specific solution to Kant's Puzzle, for instance, claiming that the notion of function is fundamental. As the reviewer recognizes, the teleosemantic project that departs from such a position may not be considered naturalist.

Before moving on, note however that I remarked that *any* solution to Kant's Problem can be the key to unknot Brentano's Problem from a naturalist standpoint. Somehow this is true if we accept both theses. This just stresses the connection between teleology and intentionality without presupposing any particular solution to Kant's Puzzle. Thus, this can be a classical etiological account or any other attempt to deal with Kant's Puzzle. By the end of the paper, I will argue that there are other sources of teleology beyond natural selection that do not face the problems that I will present later on. With these remarks in mind, I just wish to highlight that the teleosemantic project is not committed to any particular teleological position (although historically it appears to be so).

The teleosemantic project is first of all an attempt to bridge the Intentional Gap. The building blocks of the bridge are provided by biology. Mainstream teleosemantics is etiological teleosemantics. It is based on a teleological theory of trait functions—cognitive functions included. It, therefore, relies on a particular biological framework that provides a specific solution to Kant's Puzzle.

What is the solution to Kant's Puzzle endorsed by etiological teleosemantics? I am going to get into specific details in Sect. 5. For the time being, it is enough to start first by referring to its teleological basis: natural selection. The etiological theory of functions ascribes functions to traits according to the effects that they produce. Etiology can take many forms. Standard views in etiological teleosemantics are based on the Selected-Effect Theory of Functions (SETF), which is (in general) rooted in natural selection (Millikan 1989; Neander 1991). The idea is that the function of a trait is defined by those effects that such trait has had during the evolutionary process of natural selection. The function of the heart is to pump blood insofar as the effects of pumping blood make hearts being selected. Note that this theory of functions is teleological: it explains why a trait is present in nature by positing a certain purpose that it should fulfill. From this basic framework, teleosemantics extends the analysis to the cognitive domain. The core idea is that the Causal Mismatch is possible insofar as cognitive functions are selected-effect functions. Therefore, malfunctioning is possible. This means that there is room for misrepresentation. If a trait token (a cognitive function of a certain individual) does not do what it was determined by the trait type it belongs to (the evolved type), then it is possible to misrepresent the world, make errors, or produce maladaptive behaviors. Crucially, no intentionality seems to be involved in the explanation of the Causal Mismatch. This is good news because the main task is to solve Brentano's Problem in naturalistic terms. An obvious requirement for such an aim is that the explanation of intentionality does not presuppose intentionality. If we are capable of understanding the nature of intentional states in terms of evolutionary biology, the requirement is fulfilled (but see Fodor and Piattelli-Palmarini 2010).

3 Evolutionary Causation: The Causalist School

Etiological teleosemantics has been both strongly criticized and passionately defended. As a result, novel and interesting proposals have been emerging since its inception (cf. Shea 2018, Neander 2017a, Millikan 2017, for some recent proposals). Yet, most of (but not all, see, e.g., Bickhard 2003, Mossio et al. 2009) the challenges it confronts are not related to its biological foundations (its source of Natural Teleology); rather, they come from the teleosemantic analysis of representational content. My contention is that, if we accept a non-causalist, statisticalist view of natural selection (cf. Sect. 4), then SETF based on natural selection is not a fertile ground to root the causal basis of intentionality. Or, in other words, that natural selection does not provide what Brentano's Problem requires. This means, as we will see in Sect. 5 based on the Statisticalist School, that the solution to Kant's Puzzle on which etiological teleosemantics rests is unsuitable for teleosemantics' aims. To see why this is so, I need first to say something more about the solution to Kant's Puzzle that the SETF (in its evolutionary version) appeals to.

3.1 The Causal Structure of Etiological Teleosemantics

Before sketching the teleological underpinnings of etiological teleosemantics, let me first highlight what the main *explanandum* of teleological explanation is. This will help me analyze whether such teleological underpinnings really provide proper *explananda* or they need to be refurbished.

Kant's main worries were about the organizational properties of organisms. Organisms are arranged in a highly complex way that allows them to preserve and reproduce life. The complex interactions in which the many parts of an organism participate are suitable for the organism's life conditions and needs. Kant believed that this kind of complex organization, tied to organismal needs and responsible for the organism's activities, requires teleological explanations. I opt to extend the view of Kant and posit that teleological explanations deal with the adaptive dimension of organic systems. This, importantly, is not reduced to inner organizational properties but also to the external relationship that any organism bears with its environment. To use some contemporary jargon (Maturana and Varela 1980; Moreno and Mossio 2015; Kauffman 1993): we can appreciate the adaptive character of organic systems both in their *operational/organizational dimension* and in their *interactive dimension*. In this sense, teleology deploys the role of explaining how the complex organization of an organism works—through inner regulation—to fulfill its own needs and how the organism confronts its environmental conditions (through behavior or any kind of motility mediated by sensorimotor capacities) adaptively. Shortly, the moral is that *teleology is there to explain aptness*. Kant's Puzzle is to understand such explanations in causal terms. Aptness concerns the organismal capacity to sustain and reproduce one's life according to one's own needs, adequate both to

one's inner and outer conditions, or what Darwin pictured as "those exquisite adaptations of one part of the organization to another part, and to the conditions of life" (Darwin 1859).

Teleosemantics' core is to anchor intentionality in natural teleology, that is, to connect Brentano's Problem with Kant's Puzzle. While the *explanandum* of teleological explanation is the aptness of living beings, the *explanandum* of intentional explanation is one specific kind of adaptive activity that organisms perform: goal-directed behavior. This is crucial for arguing that the analysis of natural teleology incorporates the analysis of intentionality. As both riddles concern causation and goal-directed behavior is one kind of adaptive activity in organisms, the causal grounds of teleology provide a solution to the causal underpinnings of intentional behavior. As remarked, the causal basis of the SETF is natural selection. In the next subsection, I will describe the causalist view of natural selection understood in terms of populational forces.

3.2 Populational Causation

The main source of teleology on which teleosemantics rests is natural selection, that is, on explanations of aptness grounded in natural selection. Darwin was motivated by Paley's remarks of organisms' aptness and their design-like character, but he rejected Paley's answer. Henceforth, Darwinian selection became the mainstream answer for understanding the aptness of organisms, overcoming previous and competing views. As Mayr stated, "Darwin had solved Kant's great riddle" (Mayr 1991: 131). Consequently, aligned with teleosemantics' core's strategy, Darwin had provided the key to both Kant's Puzzle and Brentano's Problem. However, the causal basis for intentional causation does not come from Darwin himself but from the twentieth-century scholars working in the Modern Synthesis (henceforth, MS) framework. Intentional causation eventually rests on what came to be known as the Causalist School on natural selection, i.e., the idea that the causes of aptness lie at the populational level. We will see that, indeed, Darwin's ideas on evolutionary causes cannot be grounded in this view.

The main locus of the Causalist School on natural selection latent in the teleosemantic literature is Elliott Sober's *The Nature of Selection*.[3] The view pictured by Sober consists in looking at populations as *entities on which different causal forces act*—principally natural selection—in such a way that evolutionary biology needs *not put its hands on individual phenomena*. This signifies that *trait types*, not trait tokens, are those natural kinds on which the evolutionary causes rely, hence determining the function of the trait—what is known as *selected-effect* functions. Let's present each piece of this picture starting with the following quote:

[3] Other figures of the Causalist School are Abrams (2012), Millstein (2006), Pence (2021), Pence and Ramsey (2013), Ramsey (2016), Reisman and Forber (2005), and Stephens (2004).

> . . . the population is an entity, subject to its *own forces*, and obeying its own laws. The details concerning the individuals who are parts of this whole are pretty much irrelevant... In this important sense, population thinking involves *ignoring individuals*... (Sober 1980, p. 344, emphasis added).

Sober proposes to see that different forces act on populations through evolutionary history, in a parallel way as different forces act on objects in a Newtonian paradigm. Such forces are migration, mutation, drift, and, more importantly, insofar as it biases the course of the population in adaptive ways, natural selection. As he notes, the fact that the causes of adaptive evolution lie at the populational level opens up the possibility to ignore individuals; that is, natural selection explanations need not take into account how organisms die, live, and reproduce; they rest on the variation in fitness within groups or populations.

What are those natural kinds that provide the causes of variation on fitness that are relevant for selection explanations? In Sober's proposal, they are trait types. The fitness of a trait in a population is the core element in the *explananda* of evolutionary processes. Evolution is thus explained not by how individuals differ on fitness as a consequence of trait variation, but by how populations possessing such traits do. In Sect. 4 I will discuss whether the notion of fitness involved is a causal or a statistical one and, consequently, whether the selection process present in populational thinking involves causal or statistical explanations.

On these grounds, Sober proposes his well-known distinction between selection-for and selection-of effects (Sober 1984: 97-102). The former concern the causal role that a trait performs that makes it being selected, preserved, and spread in a population: "'Selection for' is the causal concept *par excellence.* Selection for properties causes differences in survival and reproductive success" (Sober 1984: 100, emphasis in the original). *Selection of* regards those properties of the selected trait that do not play any relevant causal role; *selection for* is causally relevant for fitness variation and selection, while the latter is not: "When there is selection for one trait and selection against another, the traits make a causal difference in survival and reproductive success" (Sober 2013: 339). Let's put an example. Imagine you have a salt shaker and you put inside two kinds of salt: one thin and white and the other thick and pink. At the time of seasoning up your dish, only the thin and white salt will go through the holes of the salt shaker, while the thick and pink one will remain inside. In this scenario, thin salt was selected for seasoning up your food because it is its being thin what made it pass through the holes; its whiteness is not a force because in the processes of selection this property did not play any relevant function, it was only selected of the population of salt grains in the salt shaker. What is relevant here is that the notions of selection-for and selection-of regard trait types, that is, whether the presence of a trait in the members of the population does or does not contribute to fitness variation.

Another important defense of the Causalist School can be found in Ernst Mayr's distinction between ultimate and proximate causation (Mayr 1961, 1974). Proximate causation belongs to the individual and ahistorical levels. Ultimate causes correspond to the evolutionary and historical levels. This also means that the difference in causes underlies a difference in the kind of question such causes are invoked to

answer. Proximate causation concerns how-questions, that is, how organisms function. Ultimate causes regard why-questions, that is, why organisms function in certain ways. Moreover, such questions are approached from different disciplines within biology. How-questions are addressed by functional biology, e.g., physiology, developmental biology, and morphology, while evolutionary biology is devoted to answering why-questions. As Mayr stated, his distinction puts some order within the many roles that each discipline in biology plays. This introduces a division of explanatory labor. The *explanandum* of each discipline can be determined according to this distinction in a way that the scope of explanatory aims is constrained and specified, concerning both what they can and should explain and what they cannot and shouldn't attempt to explain.

The distinction between ultimate/proximate causes is central to understanding the teleonomic character of living beings. Striving to stay away from the unpleasant connections that teleology has had in the history of biology, and to defend a purely mechanistic purposiveness (Mayr 1961: 1504), Mayr borrowed the term "teleonomy" from C. Pittendrigh (1958) to speak about the aptness and design-like character of individuals without non-natural connotations. In Mayr's view, a teleonomic system is any system that is the result of a program. This goes from human-made machines to living beings. As it can be appreciated, teleonomy tries to account for the aptness of living beings insofar as we can explain which are the purposes of each part of a programmed system by identifying the designed program that had built the system. The role ascribed to a system by a program, therefore, explains the capacity of the system to be intrinsically and extrinsically apt. The successfulness of the system relies both on the adequacy of the program to the system's conditions of existence and on the implementation of the program. Mayr stresses that natural selection has no teleonomic character but that individuals do. However, he can conclude that the teleonomic character of an individual is caused by ultimate causation and explained by evolution. First, he remarks that the teleonomy of individuals is a consequence of being genetically programmed (Mayr 1974: 114). Second, as natural selection belongs to the realm of evolutionary biology and it is the main ultimate cause in biology, ultimate causation and evolutionary biology tackle the teleonomy of individuals.

Even though Mayr's and Sober's views are different, there exist relevant connections concerning the SETF. The first one is that both proposals allow to speak about natural design without involving a designer, as Darwin intended. In other words, both are proposals devoted to understanding, from a naturalistic standpoint, the aptness of living beings. More importantly, in both proposals, the causes of aptness do not lie at the individual level but at the populational one. Selected-effect functions are causal at the populational level, and the teleonomic character of individuals is caused by ultimate causation coming from natural selection. Therefore, why-questions are eventually answered by evolutionary biology. Natural selection understood as a causal process at the populational level is the main tool to explain aptness. These elements constitute the process of adaptation by natural selection, where a trait "A is an *adaptation for* task T in a population P if and only if A became prevalent in P because there was *selection for* A, where the selective

advantage of A was due to the fact that A helped perform task T" (Sober 1984, p. 208, emphasis added). We can summarize the causalist position by saying that *adaptation is the key causal notion responsible for producing aptness* in living beings.

The causalist view on natural selection works for etiological teleosemantics. First, it provides the necessary causal ground. Evolutionary causation allows to understand proper cognitive functions in a naturalist way and to solve the problems underlying causation. Natural teleology is not about future stages causing current ones, but about past selection processes attributing proper functions to traits. Moreover, the mismatch between the intentional object and the world object may be explained at the level of tokens. The function of a particular representational system is to represent whatever it was designed to represent according to the type it belongs to. However, an error is understood as a deviation from evolutionary design. In other words, by attributing functions to trait types, it is possible to understand error at the token level, as a mistaken instantiation of the type it belongs to. As etiologists argue, this explanation of error does not presuppose prior intentionality. So the intentionality of the mind is rooted in a secure land. Or so it seems.

4 Evolutionary Causation: The Statisticalist School

The statisticalist view, whose foundational works are those of Walsh et al. (2002) and Matthen and Ariew (2002), as its name clearly suggests, claims that natural selection and other evolutionary processes involve statistical, not causal, explanations.

The key term in natural selection explanations is trait fitness. This, as emphasized by statisticalists (Ariew 2003; Ariew and Lewontin 2004), is different from individual fitness. Trait fitness is a property of traits belonging to a population. Individual fitness concerns individual life successes based on persistence and reproduction. Crucially, while "[t]rait fitness is the average survivability of a group of individuals possessing a type of trait" (Ariew 2003: 562), individual fitness concerns those causal processes that produce the persistence and reproduction of an individual. And here lies the difference. Trait fitness is statistically accessed, while individual fitness is causally accessed.

Populations, in populational explanations, are abstract entities. The parameters involved in such explanations are abstracted from individual-level phenomena. Populational explanations, therefore, are based on an analysis of the statistical properties of populations. The effects on populations are statistical too; they concern the distribution of trait types in a population as a function of variation on trait fitness. This constitutes the core of populational thinking the MS came to defend.

Walsh (2003, 2019) pictures two levels at which the explanation of adaptive evolution rests, the individual level and the populational level. He notes that most of the contemporary discussions around the foundations of evolutionary theory—whether MS requires no modification, an extension, or a revolution—turn around

the two-force model. It concerns the idea that beyond natural selection, as an evolutionary cause, we also have individual causes acting on evolution, in a way that discussion turns around the relevance of each level when explaining adaptive evolution. The statisticalist view is not based on the two-force model, but on the two-level one. There are no two competing levels of causes because natural selection is not an evolutionary cause. Rather, there are two types of explanations: individual, causal explanations and populational, statistical explanations. Once populational causes are removed, the statisticalist view on the causes of adaptive evolution contends, "[t]here is one level of causation; all the causes of evolution are the causes of arrival and departure (the 'struggle for life')...It is 'proximate' causes all the way down" (Walsh 2019, pp. 238, 242).

As I will note later, this provides a non-reducible and indispensable explanatory role for each explanatory level. But, before, it is important to make explicit the connection that exists, according to the Statisticalist School, between the two levels. As remarked, populations, in populational explanations, are abstract entities. This doesn't mean that populations are not composed of individuals. If we put aside the difficulty of specifying the boundaries of a population, populations could be considered sets of individuals. The abstraction of populations lies in the very explanations of populational biology—i.e., how populations are treated in the explanations done in evolutionary biology—not in the ontology of populations. And here is where trait fitness, as a statistical measure based on individual fitness, comes to the fore. Populations need not be abstract to be treated as abstract in populational explanations.

Walsh proposes to see trait fitness as an *analytic consequence* of individual fitness (Walsh 2015, 2019)—cf. Walsh (2007) for the related notion of *mere statistical effect*. The idea is that the properties of populational structures are adjudicated on the basis of the mathematical consequences of the arrival and departure of organisms (individual fitness). There are two levels of explanation, the causal-individual and the statistical-populational, where the statistical properties of the latter are a higher-order consequence/effect of the causal properties of the former. Explaining evolution, therefore, requires dealing with the consequences of individual-level causal processes at the populational level to see the changes in population structures through time. Such consequences are analytically assessed from the mathematical theory of populational biology.

As it can be appreciated, this entails a division of explanatory labor and of explanatory scope. We cannot dispense with individual-level causal explanations and population-level statistical explanations. Evolution is after all a historical and populational phenomenon. Populations evolve. Individual-level causal explanations provide individual fitness values, while population-level statistical explanations average them in terms of trait fitness variation in order to predict and explain changes in populational structures. Crucially, individual fitness is a necessary element in the explanation of evolution, yet not a sufficient one to explain evolutionary processes; we cannot explain evolution from the individual level. Trait fitness, as an abstraction of individual fitness, on the contrary, is both sufficient and necessary to explain changes in population structures insofar as it concerns those (statistical) properties of

those entities that evolve—populations (Walsh et al. 2002: 460-462). This does not mean that individuals are not relevant for evolution. Individual life span provides the causal basis of trait fitness. Without individuals, evolution has no causal roots; without populations, evolution becomes development. As Ariew defends,

> On my view evolutionary explanations are *statistical explanations of population-level phenomena* to be distinguished from "proximate" or individual level causal explanations. The result is that evolutionary explanations are indispensable even if one knows the complete causal story about how each individual in a population lived and died. In other words, evolutionary explanations are not reducible to individual-level causal explanations (Ariew 2003: 561).

Statisticalists contend that the Causalist School is not Darwinian (Walsh 2000, 2010, 2015; cf. also Godfrey-Smith (2009)). There are important differences between Darwin's proposal and the view proposed by MS. First, Darwin's insight was that the distribution of form and function deserves populational (and historical) explanations. This is common ground both in the Causalist and the Statistical Schools. Yet, the issue concerns what kinds of explanations are involved. And here lies the difference. Darwin's view is not causalist. According to him, individual causation provides the causal grounds of adaptive evolution. His notion of struggle for life came to encompass individual causation: evolution by natural selection "follow[s] inevitably from the struggle for life" (Darwin 1859) (i.e., it is a statistical effect or an analytical consequence of organisms struggling for life). The causes of life, death, and reproduction—the causes from where trait fitness is averaged—are individual causes (I will reframe Darwin's notion of "struggle" in contemporary terms in Sect. 6). Consequently, Darwinian fitness is not trait fitness but individual fitness. The view of population thinking stressed by Darwin concerns the crucial explanatory role of populational explanations, yet contrary to MS population thinking, he did not defend a causalist view of natural selection. As Walsh contends, "[t]he source of the error [in the Causalist School], I believe, lies not in the *Origin* itself but in an erroneous metaphysical picture drawn from the Modern Synthesis theory of evolution. That theory explicitly construes selection as a force acting over populations of genes" (Walsh 2000: 137). As Godfrey-Smith (2009) has shown, Darwin's definition of natural selection makes reference to struggle for life, while MS's proposals eliminate all reference to struggle for life and replace it with statistical language. The statistical underpinnings of natural selection were provided some decades after the *Origins*, as expected, by the use of statistics in the theory of natural selection (cf. Walsh (2003) on the statistical underpinnings of Fisher's Fundamental Law and the analogy with thermodynamics).

To summarize, the statisticalist view carefully distinguishes between individual fitness and trait fitness and remarks that trait fitness, which is involved in populational explanations, is statistically accessed from the variation in individual fitness. Individual fitness incorporates the proper causes of adaptive evolution. Evolution is accounted for by citing individual causes in statistical explanations: "In short, natural selection occurs only when the relative frequency of trait types changes in a population as a consequence of differences in the *average* fitness of individuals in

different trait-classes. This is what we call the statistical interpretation of natural selection" (Walsh et al. 2002: 464).

5 Challenges for Etiological Teleosemantics?

What are the consequences for etiological teleosemantics if the Statisticalist School is right? In this section, I will present two challenges that etiological teleosemantics would face if the statisticalist perspective turns out to be the correct one. In this sense, I will defend the *weak implication*: *if* the Statisticalist School is right, then etiology is challenged. Moreover, by now, I did not say anything about the connection between individual causation and natural teleology. If the solution to Kant's Puzzle rests on those processes that cause the aptness of organisms, and the causes of aptness are individual, we have to see how individual causation involves teleology and whether it can be naturalized. I leave this for Sect. 6. I will defend that the role of organismal agency in linking teleology and individual causes provides reasons for defending the *strong implication.*

The challenges presented here concern primarily the etiological theory of mainstream teleosemantics and not teleosemantics per se.[4] However, the relevance of these challenges for teleosemantics comes to the fore once we recognize that they concern the foundational grounds of teleosemantics and that, consequently, the challenges are directed to the heart of the proposal of naturalizing intentionality pursued by mainstream teleosemantics. Moreover, the literature discussed in this section comes entirely from teleosemantics, without taking into account other accounts of etiological functions that are not interested in the issue of intentionality.

5.1 Challenge 1: Functions without Causation?

The first challenge is that the solution to Brentano's Problem proposed by etiological teleosemantics fails in its biological foundations. First, note again that etiological functions are taken as selected-effect functions based on the idea of selection-for introduced by Sober: "On an etiological theory, functions are what entities were selected *for*. Mere selection *of* a trait is not enough to confer a function on it" (Neander 2017a, 132). This means that etiological functions must play the alleged causal role in evolution. Supposedly, etiological functions cause the existence of a certain trait; it is not merely about the distribution of traits within a population, as the statisticalist would argue: "Selection does more than merely distribute genotypes and phenotypes...: *by* distributing existing genotypes and phenotypes it plays a crucial

[4] I thank an anonymous review for comments on this point.

causal role in determining which new genotypes and phenotypes arise" (Neander 1995, p. 585, emphasis in original).

There is a prima facie problem here. Etiological functions cannot be selected-effect functions just because there are no such things as selected-effect functions qua populational forces. If selected-effect functions are defined on the grounds of Sober's work, then selected-effect functions are not referring to any real natural kind. Under a statisticalist view, etiological functions then, on pain of avoiding the same unhappy end, cannot be understood as selected-effect functions.

But more important is the fact that etiological teleosemantics fails to solve Brentano's Problem. This failure lies in the very teleosemantics' core outlined before. As explained, the core idea of any teleological theory of content is to solve Brentano's Problem by solving Kant's Puzzle. Unfortunately, if the statisticalist reading is right, the biological solution to Kant's Puzzle on which etiological teleosemantics rests is wrong. The statisticalist's thesis is universal. All causes of evolution lie at the ontogenetic level. Thus, there is no way that causalists can solve Kant's Puzzle, principally because, under the statisticalist reading, natural selection has no causal weapons to explain the Causal Asymmetry of teleological explanations. Since populational explanations of natural selection processes are not causal, they are not suitable to ground the causal basis of teleology. In this sense, in etiological teleosemantics, the teleosemantics' core itself fails: Brentano's Problem cannot be approached if Kant's Puzzle is not solved first.

As explained, the trouble begins when we understand that selected-effect functions are based on trait fitness not on individual fitness. Thus, they are not causal, and, as a consequence, neither are etiological functions. Rather, we can claim that etiological functions could be understood as statistical functions based on the change in populational structure as a consequence of the statistical properties of the population—which are a consequence of individual fitness. *I think that this is the proper reading of etiological functions: as statistical functions.* So, one may ask, why etiological teleosemantics doesn't work under a statisticalist view of etiological functions?

The main reason is that statistical functions cannot provide the causal grounds for teleological functions; thus, they fail in any attempt to deal with Causal Asymmetry. But there is another important point to highlight which concerns the explanatory role of functions. As Garson and Papineau, defending etiological functions, describe it:

> First, functions are explanatory. One peculiar feature of functions is that, when biologists attribute a function to a trait, they are often trying to give *a causal explanation for why that trait exists*. One virtue of the selected effects theory is that it makes sense of this explanatory aspect of functions (Garson and Papineau 2019: 4, emphasis added).

This is an ontological issue. As these authors remark, the teleological function of a trait must explain why such a trait exists. A teleological function is connected with the adaptiveness of a trait—the adaptiveness of a trait being due to having such function, that is, with the causal role of such function that makes it part of nature. As they highlight, such explanation must be causal, concerning the process that made this trait part of nature. Accordingly, under a causalist view, as Garson and Papineau

argue, etiological functions are suitable for this task. In etiological teleosemantics, the proper functions of a representational system concern those causal connections that have produced a goal-directed behavior that made a certain trait type being selected. The problem now is quite expected. Etiological functions, under its statisticalist definition, do not provide a causal explanation for the existence of traits. Thus, they cannot accomplish the explanatory role that functions have. Etiological functions, qua statistical functions of trait types, do not explain the existence of a trait in causal terms.

Bickhard (2003) has argued that etiological functions are causal epiphenomenal at the individual level. This is so because etiological functions do not concern how individual systems operate; they are not based on the causal processes within a particular token. This is true independently of one's commitment to statisticalism. The solution of error promoted by etiological teleosemantics rests on the possibility that the causal processes in an individual perform a function that deviates from the functions specified by the evolved trait type; i.e., error is a mismatch between trait tokens and trait types. However, if we accept the statisticalist view, this section concludes that etiological functions are also causal epiphenomenal at the evolutionary level. Statistical, etiological functions are the statistical effect of individual causes. Etiological functions, therefore, if statisticalism is right, are causal epiphenomenal both at the ontogenetic and the phylogenetic levels.

5.2 Challenge 2: Statistical Norms for Nonstatistical Explanations?

This second challenge is a direct consequence of the first one, insofar as etiological functions provide the normative dimensions of content needed to solve the Causal Mismatch. Once again, this section is based on the assumption that statisticalism is right. From this standpoint, the problem is, roughly, that the kind of norm provided by etiological functions is not the kind of norm needed for teleosemantics. My argument is structured as follows:

Premise: Statisticalism is right.

1. Teleosemantic norms cannot be statistical.
2. But etiological functions under the statistical view are statistical functions.
3. Then, etiological functions provide statistical norms.
4. Therefore, etiological norms are not teleosemantic norms—i.e., they are not the norms that a teleosemantic project can appeal to. Etiological functions are not the kind of function adequate for teleosemantics.

Conclusion: If statisticalism is right, etiological norms are not teleosemantics norms.

Surely, the critical step is 1. There are two ways of arguing for it. The first one is quite straightforward: to take a look at the literature on etiological teleosemantics to

see why etiological norms cannot be statistical. The second way is by providing an argument independent of any specific literature. Let's start by the first one.

The first way to defend point 1 is by noting that etiologists themselves also defend it. A central requisite for a theory of norms is that norms cannot be statistical. This is highlighted by different scholars; thus, Neander: "It might help to note that the normativity of biological functions is neither simply evaluative or statistical" (Neander 1995: 111). A theory of proper functions must give more than statistical generalizations, but "[t]he description of the normal system as the system that functions 'as designed' is thus not merely a generalization but a useful generalization in ways that surpass mere statistical generalization" (Neander 2017b, p. 1161).

The clearest example in the literature is found in the work of Ruth Millikan. She even promotes a typographical distinction in her use of "Normal" instead of "normal" (for instance, in Millikan (1984, chs. 1 and 2) and Millikan (2017, ch. 6)). She "...capitalize[s] *Normal* —to distinguish it from *normal* in the sense of *average*" (Millikan 1984, p. 34, italics in the original). The difference, therefore, entails that Normal denotes a proper function, while normal is just a mere average, statistical distribution in a population. "Proper functions do not concern norms in any evaluative or prescriptive sense. They do not concern norms in a statistical sense either. On the contrary, there are many items that usually fail to perform their proper functions" (Millikan 2000: 88).

The conclusion from these quotes is that the norms for teleosemantics cannot be statistical. Etiological functions, under a causalist reading, are not statistical but causal. They do more than just point out the typicality or high average of traits, and they provide a causal criterion to determine when a trait token is functioning properly and when it is not.

This point is usually considered an important advantage of etiological functions over Cummins-functions (Cummins 1975). As Cummins-functions are not teleological functions, the demarcation between proper and improper cannot be based on the natural purposes or the goals of a trait; rather, at most, Cummins-functions can only provide a statistical criterion for function/malfunction demarcations:

> By contrast [with the etiologists], all that the systems account [Cummins Functions] can offer is a statistical criterion: in *most* systems of a certain kind this kind of trait does F, so here the trait is malfunctioning in not doing F. By contrast with the etiological analysis, this statistical systems account seems to lack any normative content: it doesn't seem to show that a trait in any sense *ought* to be doing F; it just says it *isn't* doing F, and so is statistically unusual, but nothing more (Macdonald and Papineau 2006; pp.11-12, emphasis in original).

However, there is a way to refute my first defense of point 1. One could refute my argument as follows: the rejection of statistical norms by etiologists aims at demarcating the sense of normal—as high average or typical—from proper normativity. That is, that a trait is frequent or usual does not mean that it was selected for during evolution.[5] As Neander explains in connection to pandemic scenarios:

[5] Note that Neander's example concerns the current frequency of a trait achieving or not its function. However, this scenario can also be presented in relation to past, historical frequencies. Let's appeal

> there is no incoherence in the idea that functional impairment could become typical in a population for a time, in a pandemic or due to an environmental disaster. The relevant function-dysfunction distinction does not seem to be simply the typical-atypical or expected-unexpected activity distinction. This much is fairly uncontroversial (Neander 2017b: 1152).

Let's call "typical-norms" these kinds of norms based on averages. The central point to rebut my critique is that etiological norms, even if statistical, are not based on the notion of average or typicality. Etiological functions and etiological norms are based on trait fitness. The process of natural selection is explained by the change in populational structure due to its statistical properties. But this doesn't mean that trait fitness and populational change entail that the selected trait is typical or normal. A trait with a low frequency could have been selected notwithstanding. Trait fitness is not related to average. The attack to my critique does not deny point 2, but it does reject point 1. *Even if* etiological norms are statistical norms, etiological norms are not typical-norms. Point 1* should be: Teleosemantic norms cannot be typical-norms. So my critique does not work. Unless we have more reasons to defend point 1.

To answer this counterargument, there is a second way to defend point 1 besides any specific literature on teleosemantics. I believe that there are good reasons to defend that natural norms for teleosemantics cannot be statistical (i.e., to defend 1). But first, let's consider the following point. In some sense, the presented counterargument is sound. Trait fitness is not just a matter of high average. Yet, etiological theories are based on the bias that natural selection introduces that results in a trait type being selected, preserved, and spread. Although trait fitness is not just high-average, trait fitness is related to the increment of the frequency of a trait in a population. The notion of selection for, both under the causalist and the statisticalist readings, concerns the consequences for a population when its individuals possess the selected trait. Such consequences do concern the increment on average in the population. That a trait has been selected is related to what contributions such a trait had provided that make it more frequent and reliable in a population; similarly, if a trait is "selected against" (Sober 1984), it means that natural selection biased it towards a low frequency in the population. In this sense, trait fitness is related to an increment in average. Besides this preliminary point, there are further and more relevant reasons to discard the alleged refutation.

The central reason lies in the very idea of appealing to selected-effect functions—qua populational forces—to solve the Causal Mismatch. The etiological solution to the problem of misrepresentation needs to consider selected-effect functions as causal dispositions. The central idea of this solution lies in understanding error as

---

to the case of sperm, used by Millikan. Rightly, she points out that the current frequency of a trait function does not lead us directly to its proper function—sperms perform their proper function at a very low frequency. However, this argument applies not only to current frequencies but also to historical frequencies (its frequency during selection processes): it is not necessary that in selection processes a trait must have a high frequency to be selected for. Sperms could perform their proper function infrequently and nonetheless be selected to perform such function. I thank an anonymous review for comments on this point.

a mismatch between the evolved trait type and the individual trait token. The normativity of content is not attributed at the ontogenetic level but at the phylogenetic one. Crucially, the processes of normative attribution must be causal. The explanatory target is to understand how representation and reference can be causally linked even if the possibility of error is taken into account. This is what makes teleosemantics a naturalist project. To account for the Causal Mismatch, one must explain how the normative content of a representation is attributed in a way that the reference of such representation is understood in causal terms. Etiology, under a causal reading, can provide such kind of causal explanations, insofar as what specifies the normativity of content is the causal role that a trait type has had that made it part of the population. But, if the normativity of content is explained statistically, the connection between a representation and its reference would lack the causal grounds needed for the naturalization of intentionality.

Here we can appreciate the connection with challenge 1. Teleosemantics' core works for the naturalization of intentionality insofar as it can account for intentional causation from a legitimate scientific viewpoint. The strategy to appeal to etiological functions makes it possible to root the normativity of content in solid causal grounds. However, insofar as such causal grounds are not the alleged ones by etiologists (challenge 1), the causal roots of normativity are cut. Hence, we have a crucial reason to defend point 1 (that teleosemantic norms cannot be statistical). The normativity of content that specifies the extension of any representation must be assessed in causal terms in order to provide a naturalist view of intentionality. If my considerations are correct, point 1 stands, and consequently, my critique concludes that, under a statisticalist view, etiological norms, and therefore selected-effect functions too, are not suitable for a naturalistic solution to Brentano's Problem.

To conclude, let's appreciate a defense of point 1—besides the issue of typical-norms—in the teleosemantic literature, particularly in Millikan's emphasis on "Normal explanations." Her emphasis on Normal (capitalized) traits involves not merely to separate them from typical-norms. It also stresses the kinds of natural norms that teleosemantics needs and what kind of explanation specifies the normativity of content. Normal explanations (of a representation) concern the causal role of a trait that had assigned (past tense) a normative content to the representation. Note that a Normal explanation need not be identified with any specific theory of teleological function. If we look at it with etiological (Millikan's) lenses, such causal role is attributed to *trait types* and concerns the historical conditions that made a trait type being selected. Accordingly, the historical process of natural selection is what attributes normative content to representational systems. However, here we find again the aforementioned problem: under a statisticalist view, etiological functions do not provide Normal explanations; causal roles cannot be attributed to trait types. As a consequence, Normal explanations, if causal, are not etiological explanations of trait function.

6 Agential Teleosemantics: A New Solution to Kant's Puzzle, a New Solution to Brentano's Problem

6.1 Agential Teleology

My proposal till now states that if we adopt a statisticalist view of natural selection, etiological teleosemantics has a foundational problem in its solution to Brentano's Problem. In this last section, I intend two things. First, I will argue that teleosemantics' core is still a valid and interesting path towards the naturalization of intentionality. However, to the extent that the condition of etiological teleosemantics highlighted here is present in its teleological framework, it is necessary to propose an alternative view of natural teleology in order to reframe teleosemantics on proper causal grounds. In other words, we need to propose an alternative solution to Kant's Puzzle to anchor a new answer to Brentano's Problem. The ideas presented here are just a sketch that requires further elaboration, and thus I do not expect to solve everything but just to propose a different strategy for teleosemantics in the context of the causalism vs. statisticalism debate on the causal basis of natural selection.

Secondly, I will provide some reasons why the statisticalist view is correct. Therefore, while till now I just defended the *weak implication* (i.e., an analysis of the implications for etiological teleosemantics *if* we adopt a statisticalist viewpoint), to point out the main reasons why statisticalism is true entails a defense of the *strong implication.* As expected, both points are connected: a defense of statisticalism consists in pointing out an alternative locus of evolutionary causes and, consequently, to link such causes with the causal explanation of aptness—i.e., the grounds of natural teleology. In other words, the defense of an individual-level account of natural teleology operates as an argument in favor of the statisticalist view, insofar as it means a defense of the causal grounds of the explanation of aptness at the individual level, as statisticalists support. So, now it is necessary to specify what is the locus of evolutionary causes and how it connects with natural teleology and, consequently, with the causes of aptness.

"The only genuine forces going on in evolution are those taking place at the level of individuals" (Walsh et al. 2002: 453), so the Statisticalist School argues. Individual causes of evolutionary processes concern those that specify the fitness of an individual—its persistence and reproduction, that is, the life, death, and reproduction of an organism. This includes what Darwin referred to as the *struggle for life*: organisms pursuing successful life conditions in such a way that their fitness increases. As it can be appreciated, struggling is an action and an activity that an individual organism performs. Although struggling has a connotation of competitiveness, we can reinterpret it in contemporary terms by claiming that aptness is caused by the *adaptivity* of organisms as *agents* (Sultan et al. 2022; Walsh 2015).

The core idea is that the organism itself, not its genes, defines and regulates its ontogenetic trajectory. Such trajectory is adaptive insofar as the organism is constantly adjusting it according to both its inner and outer circumstances. Agentivity

hinges on the capacity of an organism to self-regulate its life conditions in adaptive ways in order to endure and reproduce. This view posits organismal activity not as the execution of a genetic program but as a goal-oriented process carried out by the organism itself towards an adaptive phenotypic state. The goal-directedness of organismal activities stresses their adaptive dimension. That's why Walsh (2007, p. 195) states that, "Evolution is adaptive because ontogeny is adaptive." Therefore, "[t]hose exquisite adaptations of one part of the organization to another part, and to the conditions of life" (Darwin 1859) are not caused by genetic programs nor by evolutionary design, but by what agents do during development to stay adapted.

As it can be appreciated, the causes of evolution are proximate causes. Teleology, therefore, lies in the activity of organisms during their life span that determines their individual fitness. Importantly, while Mayr posits that functional biology is devoted to analyzing proximate causes, from a contemporary perspective we can locate proximate causes of aptness in two broad frameworks devoted to the study of individuals: organismic biology and developmental biology. The term "developmental biology," which is concerned with the temporal dimension of individual development, today embraces many frameworks that have challenged certain core tenets of MS, such as evo-devo, eco-devo, developmental systems theory, ontogenetic theories of inheritance, and developmental psychobiology. Although there is not a unified view, in all cases the task is to transcend the gene-centered view of development and evolution and provide a theory of development where the organism is the proper unit of development. Organismal biology has a philosophical connotation, but it refers to the physiological level: how parts interact in order to produce a certain outcome. Note that this is usually understood in mechanistic terms; thus, it seems that there is no place for teleology. The emphasis on organismal biology is because organismal biology does posit organisms as teleological agents while still relying on the physiology and functioning thereof. The connection between these areas requires further exploration. If the revival of teleology at the individual level involves two different temporal scales—the organismic and the developmental one—we need to unify them in order to propose a coherent picture of natural teleology. Let's briefly point out some core processes of both frameworks.

In recent decades a number of mechanisms have been put forward with the aim of constructing a view of ontogeny as a context-sensitive, multi-causal, and complex process. This involves different areas of research, such as niche construction theory (Odling-Smee et al. 2003; Stotz 2017), developmental studies on phenotypic plasticity (Bateson and Gluckman 2011; West-Eberhard 2003), the study of norms of reaction (Schlichting and Pigliucci 1998; Sultan 2015), ecological developmental biology (Gilbert and Epel 2015; Lewontin 2001), epigenetic systems of inheritance (Jablonka and Lamb 2005), developmental approaches to homology (Wagner 2014), molecular epigenetics (Griffiths and Stotz 2013; Keller 2002; Moss 2003), psychobiological development (Gottlieb 1997; Keller 2010; Michel and Moore 1995), and embryology (Amundson 2005; Robert 2006). In all cases, we have organisms understood as subjects self-regulating their development in an adaptive way according to their living conditions, not as the implementation of an evolved developmental program. Within organismic biology, we also find many interesting

phenomena that have been investigated during the last decades, such as the self-organizational capacities of organisms (Camazine et al. 2003; Goodwin 2001; Kauffman 1993; Mitchell 2009; Müller and Newman 2003), global dynamics studies by systems theorists (Bertalanffy 1969; Boogerd 2007; Noble 2008), the autopoietic (Di Paolo 2005; Maturana and Varela 1980) and cybernetic (Ashby 1991) properties of organisms, and the autonomy of organic systems (Barandiaran et al. 2009; Moreno and Mossio 2015), among others. Here, we also find organisms doing things to stay adaptive, in such a way that their physiology cannot be understood without taking into account the directedness of these processes towards an adaptive way of being alive. This has led many scholars, both within the organismic and the developmental frameworks, to defend the agentive view of the organism and its teleological underpinnings.

Precisely, these recent emphases on individual causation give us reasons to support the statisticalist view and, consequently, argue for the *strong implication*. As highlighted in the introduction, I connected the position adopted concerning the role of organisms in evolution with different sides of the statisticalism vs. causalism debate. If organisms are irrelevant in explaining evolution—as Sober argued—it seems that causes cannot lie at the individual level; but if organismal agency is the locus of the solution to Kant's Puzzle, individual causation provides the proper causes of evolution, as statisticalists claim.[6] The rise of developmental and organismal biology has come to challenge the evolutionary view pictured by MS. Recent views in biology, and their theoretical and experimental backup, promote reasons for defending the *strong implication*: statisticalism is true, and, therefore, teleosemantics must be challenged and reframed.

From what I have been presenting, I conclude that the causes of evolution are explained by the teleological dimension of organisms as agents. Let's call this Agential Teleology. In the vein of the picture developed here, Agential Teleology explains the aptness of living beings by stressing the adaptivity of organic agents. Adaptivity not adaptation, agents not populations, provides the causal bases of natural selection.

6.2 *Towards Agential Teleosemantics*

Agential Teleosemantics is rooted in Agential Teleology. This means that the proper function of any representational system is established during individual ontogeny. While etiologists claim that a representational system must represent whatever it was

[6] Certainly, one can take a midterm position and argue for both individual and populational causes of evolution. As Walsh noted (2003, 2019), this view defends a two-force model (cf. Sect. 4): evolutionary forces come from two different sources. However, as it was remarked above, this view is not endorsed by statisticalists insofar as it blurs the nature of populational explanations based on trait fitness. There are not two forces but two different levels of explanation (cf. Walsh (2019) for details).

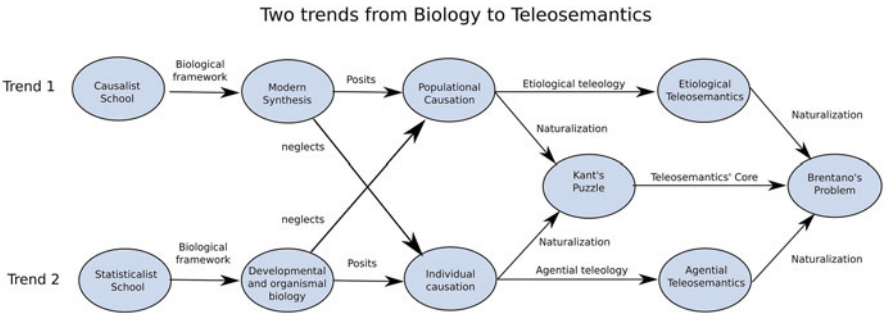


Fig. 2 Two trends from biology to teleosemantics. Trend 1 is defended by the Causalist School based on the MS framework. The MS posits populational causes to provide a naturalist solution to Kant's Puzzle while neglecting individual causation as an evolutionary force. Populational causation is the central foundational element of etiological teleosemantics and therefore the grounds for the naturalist solution to Brentano's Problem. Trend 2 is defended by the Statistical School based on the biological framework recently pursued by developmental and organismal biology. Such a framework posits individual causes to provide a naturalist solution to Kant's Puzzle while neglecting populational causation. Individual causation is the central foundational element of agential teleosemantics and therefore the grounds for the naturalist solution to Brentano's Problem

selected for during natural evolution, in Agential Teleosemantics the possibility of error and the explanation of Causal Mismatch hinge on what a system must represent according to those proper functions developed during ontogeny and performing a particular causal role in any organismal activity. Insofar as the developmental and organismal dimensions of organisms have a teleological flavor, the teleosemantics' core is preserved from etiological teleosemantics. Agential Teleology posits teleological functions, hence a normative valuation on traits according to their individual history in organismal development. This, therefore, accounts for the problem of misrepresentation. As Agential Teleology does not involve prior intentionality, it seems that this new solution to Kant's Puzzle provides an alternative naturalistic position for dealing with Brentano's Problem. Figure 2 locates this proposal in biology by stressing how different biological frameworks provide alternative teleosemantic projects.

Certainly, I didn't say anything specific about the core ingredients of teleosemantics: its *teleo* side and its *semantic* side. I did not propose a theory of functions based on the adaptive agency of organisms (but see the organizational view of function presented by Mossio et al. (2009) or the view of function defended by Walsh (2014) in relation to plasticity). Nor I said anything about the nature of representational systems in relation to content determination—for instance, whether Agential Teleosemantics fits with an input-based or output-based account. This is for further work. By now, I limit myself to sketch a different proposal motivated by the discussion about the nature of natural selection.

The view outlined here has many points in contact with different proposals within teleosemantics, particularly with Dretske's account (Dretske 1981, Dretske 1988).

Even though he did not put up his view with the focus on the agential capacities of organisms recently defended in theoretical biology, nor he directly took part in the discussion around evolutionary causation, he did defend that the processes responsible for attributing functions to representations (more specifically to type 3 systems of representation; Dretske 1988: ch. 3) reside in individual development. Moreover, other contemporary proposals, while not anchored to current debates on evolutionary causation, pursue a similar strategy as Agential Teleosemantics, such as Bickhard's interactivism (Bickhard 2000, 2009), de Prado Salas (2018) on reproduction, or Schroeder's (2004) emphasis on cybernetic properties of cognition. It is expected, then, that *ontogenetic selected functions* do have many points of contact with Agential Teleosemantics. This includes Shea's recent emphasis on the role of persistence in function determination (Shea 2018), Millikan's insight on operational condition and external inheritance systems (Millikan 1984), or Garson and Papineau's proposal on novel content (Garson and Papineau 2019). These remarks help me to constrain the scope of my critics. My proposal here is foundational, and it touches on some parts of the teleosemantic project, which, although central, does not purport a rejection of teleosemantics at all. On the contrary, my proposal is totally aligned with teleosemantics' aims. Having said so, I do believe that many proposals on ontogenetic function fit with Agential Teleosemantics. Moreover, I did not discuss the specific proposals concerning the content determination and the variant of teleosemantics accounts (such as informational teleosemantics, consumer-based accounts) and their possible connections with Agential Teleosemantics (but see the last paragraph).

Let's conclude with some programmatic open questions. Most of the proposals of teleosemantics are devoted to explain the cognitive capacities of animals—bee dance, frogs catching flies, human beliefs, and so on. This is not the case, for instance, with Nicholas Shea's account. Although his 2018 book is devoted to cognitive science, his previous work on teleological functions was not restricted to the cognitive level but pursued a teleosemantic account applied to all living beings. I am not going to discuss his proposal here (cf. Griffiths 2013; Author 2022). I just want to stress that Agential Teleosemantics need not concern animal cognition exclusively. In this sense, it represents a philosophical project both for cognitive science and for theoretical biology. This means that the two core ingredients of Agential Teleosemantics are latent in all individuals. The first one, concerning the teleological side of teleosemantics, is quite simple to defend, together with etiological theories of functions, insofar as it comes from biology itself. Thus, if Agential Teleosemantics is based on the teleological dimension of agents, and all living beings are agents, therefore the teleological notion of function defended here is present beyond animal cognition.

The second ingredient, concerning semantics, promotes an open question for further analysis. Should we extend semantics beyond animal cognition? This is a difficult issue insofar as semantics is usually employed as an umbrella term including different phenomena. Shea accepts it and suggests an informational teleosemantics beyond animal cognition. I propose that the possibility to extend semantics beyond animal cognition concerns the use of informational talk in biology, particularly in

developmental and organismic biology, and to the extent that it is in these areas where Agential Teleology is based on. If we set aside the etiological functions defended by Shea, and endorse a view on function based on Agential Teleology, there are good reasons to extend Agential Teleosemantics beyond animal cognition. In a nutshell, informational processes are central to the teleological character of agents. As remarked, Agential Teleology concerns the adaptivity of each organism to its life conditions. This is the common ground that connects different phenomena within developmental and organismic biology. In all these cases, organisms adaptively confront their living conditions in order to persist and reproduce. Crucially, *if such agential capacity is adaptively directed*, that is, directed towards an adaptive way of being a living system, then organismal responses, through regulation, niche construction, self-organization, plasticity, and so on, *must be sensitive to its life conditions* (Author 2021). Here information enters the scene. Information is central to explain how an agent's activity—from cell development and the physiology of organs to motility and behavior—is sensitive to both its inner and outer conditions in such a way that the performed activity is adaptive to such conditions. In sum, adaptive agents are informed; sensitiveness is a prerequisite for Agential Teleology. In this sense, a research question within Agential Teleosemantics concerns the possible unification of informational talk between developmental and organismic biology, for instance, as it is suggested by developmental system theorists (Oyama 1985; Griffiths 2016, 2017; Griffiths et al. 2015; Stotz 2006, 2019; Calcott et al. 2020; Griffiths and Stotz 2013), and cognitive science, such as the proposals of Shea (2018), Neander (2017a), Dretske (1981, 1988), or even Fodor (1998).

Acknowledgments I thank the useful comments and suggestions of an anonymous reviewer. I am grateful to Sergio Balari for the discussion of the ideas presented here and for his help to make the language of this paper closer to readable English. This work has received the support of the Spanish Government through grant FFI2017-87699-P and the National Agency of Investigation and Innovation (Uruguay) through grant POS_EXT_2018_1_154759.

References

Abrams M (2012) Measured, modeled, and causal conceptions of fitness. Front Genet 3:196
Amundson R (2005) The changing role of the embryo in evolutionary thought. Cambridge University Press
Ariew A (2003) Ernst Mayr's' ultimate/proximate'distinction reconsidered and reconstructed. Biol Philos 18(4):553–565
Ariew A, Lewontin R (2004) The confusions of fitness. Br J Philos Sci 55(2):347–363
Ashby R (1991) Principles of the self-organizing system. In: Klir GJ (ed) Facets of systems science. Springer, Boston, MA, pp 521–536
Author (2021) Biosemiotics at the bridge of eco-devo and representational theories of mind. Ital J Philos Language 15(2):59–92
Author (2022) Agential Teleosemantics. PhD Thesis. Autonomous Universtiy of Barcelona.
Barandiaran X, Di Paolo E, Rohde M (2009) Defining agency: individuality, normativity, asymmetry, and Spatio-temporality in action. Adapt Behav 17(5):367–386

Bateson P, Gluckman P (2011) Plasticity, robustness, development and evolution. Cambridge University Press
Bertalanffy LV (1969) General system theory: foundations, development, applications. George Braziller, New York
Bickhard M (2000) Information and representation in autonomous agents. Cogn Syst Res 1(2): 65–75
Bickhard M (2003) The biological emergence of representation. In: Brown T, Smith L (eds) Reductionism and the development of knowledge. Psychology Press, pp 115–142
Bickhard M (2009) Interactivism: a manifesto. New Ideas Psychol 27(1):85–95
Boogerd F (2007) Systems biology: philosophical foundations. Elsevier, Amsterdam, Boston
Bromberger S (1966) Why-questions. In: Colodny R (ed) Mind and cosmos. University of Pittsburgh Press, pp 86–111
Calcott B, Pocheville A, Griffiths P (2020) Signals that make a difference. Br J Philos Sci 71(1): 233–258
Camazine S, Deneubourg J, Franks NR et al (eds) (2003) Self-organization in biological systems. Princeton University Press, Princeton
Cummins R (1975) Functional analysis. J Philos 72(20):741
Darwin C (1859) On the origin of species: by means of natural selection or the preservation of favoured races in the struggle for life. John Murray, London
Di Paolo EA (2005) Autopoiesis, adaptivity, teleology, agency. Phenomenol Cogn Sci 4(4): 429–452
Dretske F (1981) Knowledge and the flow of information. MIT Press, Cambridge, MA
Dretske F (1988) Explaining behavior. MIT Press, Cambridge, MA
Fodor J (1998) Concepts. Where cognitive science went wrong. OUP, Oxford
Fodor J, Piattelli-Palmarini M (2010) What Darwin got wrong. Profile, London
Garson J, Papineau D (2019) Teleosemantics, selection and novel contents. Biol Philos 34(3):36
Gilbert S, Epel D (2015) Ecological developmental biology: the environmental regulation of development, health, and evolution. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts
Godfrey-Smith P (2009) Darwinian populations and natural selection. Oxford University Press, Oxford
Goodwin B (2001) How the leopard changed its spots. Princeton University Press, Princeton
Gottlieb G (1997) Synthesizing nature-nurture: prenatal roots of instinctive behavior. Taylor Francis Inc.
Griffiths P (2013) Lehrman's dictum: information and explanation in developmental biology. Dev Psychobiol 55(1):22–32
Griffiths P (2016) Proximate and ultimate information in biology. In: Couch M, Pfeifer J (eds) The philosophy of Philip Kitcher. Oxford University Press, pp 74–97
Griffiths P (2017) Genetic, epigenetic and exogenetic information in development and evolution. Interface Focus 7(5):20160152
Griffiths P, Pocheville A, Calcott B et al (2015) Measuring causal specificity. Philos Sci 82(4): 529–555
Griffiths P, Stotz K (2013) Genetics and philosophy. Cambridge University Press, Cambridge
Hamburger V (1980) Embryology and the modern synthesis in evolutionary biology. In: Mayr E, Provine W (eds) The evolutionary synthesis. Perspectives on the unification of biology. Harvard University Press, pp 97–112
Jablonka E, Lamb M (2005) Evolution in four dimensions : genetic, epigenetic, behavioral, and symbolic variation in the history of life. MIT Press, Cambridge, MA
Kauffman S (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, New York
Keller EF (2002) The century of the gene. Harvard University Press, Cambridge, Mass. London
Keller EF (2010) The mirage of a space between nature and nurture. Duke University Press, Durhnam, NC

Lewontin R (2001) The triple helix: gene, organism, and environment. Harvard University Press, Harvard
Macdonald G, Papineau D (2006) Introduction: prospects and problems for teleosemantics. In: Macdonald G, Papineau D (eds) Teleosemantics. Clarendon Press, pp 1–22
Matthen M, Ariew A (2002) Two ways of thinking about fitness and natural selection. J Philos 99(2):55–83
Maturana H, Varela F (1980) Autopoiesis and cognition. Springer, Netherlands
Mayr E (1961) Cause and effect in biology: kinds of causes, predictability, and teleology are viewed by a practicing biologist. Science 134(3489):1501–1506
Mayr E (1974) Teleological and teleonomic, a new analysis. In: Cohen R, Wartofsky M (eds) A portrait of twenty-five years, Boston studies in the philosophy of science. Springer, Dordrecht, pp 133–159
Mayr E (1991) The ideological resistance to Darwin's theory of natural selection. Proc Am Philos Soc 135(2):123–139
Michel G, Moore C (1995) Developmental psychobiology: an interdisciplinary science. MIT Press, Cambridge, MA
Millikan R (1984) Language, thought, and other biological categories: new foundations for realism. MIT Press, Cambridge, MA
Millikan R (1989) In Defense of proper functions. Philos Sci 56(2):288–302
Millikan R (2000) Naturalizing intentionality. Philosophy Documentation Center 9(83):83–90
Millikan R (2017) Beyond concepts: unicepts, language, and natural information. Oxford University Press, Oxford
Millstein R (2006) Natural selection as a population-level causal process. Br J Philos Sci 57(4): 627–653
Mitchell M (2009) Complexity: a guided tour. Oxford University Press, Oxford
Moreno A, Mossio M (2015) Biological Autonomy. Springer, Netherlands
Moss L (2003) What genes can't do. MIT Press, Cambridge, MA
Mossio M, Saborido C, Moreno A (2009) An organizational account of biological functions. Br J Philos Sci 60(4):813–841
Müller G, Newman S (eds) (2003) Origination of organismal form: beyond the gene in developmental and evolutionary biology. MIT Press, Cambridge, MA
Neander K (1991) Functions as selected effects: the conceptual Analyst's Defense. Philos Sci 58(2): 168–184
Neander K (1995) Misrepresenting and malfunctioning. Philos Stud 79(2):109–141
Neander K (2017a) A mark of the mental. MIT Press Ltd, Cambridge, MA
Neander K (2017b) Functional analysis and the species design. Synthese 194(4):1147–1168
Noble D (2008) The music of life: biology beyond genes. Oxford University Press, Oxford
Odling-Smee J, Laland K, Feldman M (2003) Niche construction: the neglected process in evolution. Princeton University Press, Princeton
Oyama S (1985) The ontogeny of information, 2nd edn. Duke University Press, Durhnam, NC
Pence C (2021) The causal structure of natural selection. Cambridge University Press, Cambridge
Pence C, Ramsey G (2013) A new foundation for the propensity interpretation of fitness. Br J Philos Sci 64(4):851–881
Pittendrigh CS (1958) Adaptation, natural selection, and behavior. Behav Evol 390:416
Potochnik A (2017) Idealization and the aims of science. University of Chicago Press, Chicago
de Prado Salas JG (2018) Whose purposes? Biol Teleol Intentionality Synthese 195(10):4507–4524
Ramsey G (2016) The causal structure of evolutionary theory. Australas J Philos 94(3):421–434
Reisman K, Forber P (2005) Manipulation and the causes of evolution. Philos Sci 72(5):1113–1123
Robert JS (2006) Embryology, epigenesis and evolution. Cambridge University Press, Cambridge
Schlichting CD, Pigliucci M (1998) Phenotypic evolution: a reaction norm perspective. Sinauer Associates Incorporated
Schroeder T (2004) New norms for teleosemantics. In: Clapin H, Staines P, Slezak P (eds) Representation in mind. Elsevier, pp 91–106

Shea N (2018) Representation in cognitive science. Oxford University Press, NY
Sober E (1980) Evolution, population thinking, and essentialism. Philos Sci 47(3):350–383
Sober E (1984) The nature of selection: evolutionary theory in philosophical focus. University of Chicago Press, Chicago
Sober E (2013) Trait fitness is not a propensity, but fitness variation is. Stud Hist Philos Sci Part C: Stud Hist Philos Biol Biomed Sci 44(3):336–341
Stephens C (2004) Selection, drift, and the "forces" of evolution. Philos Sci 71(4):550–570
Stotz K (2006) With 'genes' like that, who needs an environment? Postgenomics's argument for the 'ontogeny of information'. Philos Sci 73(5):905–917
Stotz K (2017) Why developmental niche construction is not selective niche construction: and why it matters. Interface Focus 7(5):20160157
Stotz K (2019) Biological information in developmental and evolutionary systems. In: Uller T, Laland K (eds) Evolutionary causation: biological and philosophical reflections. The MIT Press, Cambridge, pp 323–344
Sultan S (2015) Organism and environment: ecological development, niche construction, and adaption. Oxford University Press, New York
Sultan S, Moczek AP, Walsh D (2022) Bridging the explanatory gaps: what can we learn from a biological agency perspective? BioEssays 2100185:1–14
Wagner G (2014) Homology, genes, and evolutionary innovation. Princeton University Press, Princeton
Walsh D (2000) Chasing shadows: natural selection and adaptation. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 31(1):135–153
Walsh D (2003) Fit and diversity: explaining adaptive evolution. Philos Sci 70(2):280–301
Walsh D (2007) The pomp of superfluous causes: the interpretation of evolutionary theory. Philos Sci 74(3):281–303
Walsh D (2010) Two neo-darwinisms. Hist Philos Life Sci 32(2):317–339
Walsh D (2014) Function and teleology. In: Thompson P, Walsh D (eds) Evolutionary biology. Cambridge University Press, Cambridge, pp 193–216
Walsh D (2015) Organisms, agency, and evolution. Cambridge University Press, Cambridge
Walsh D (2019) The paradox of population thinking: first order causes and higher order effects. In: Uller T, Laland K (eds) Evolutionary causation: biological and philosophical reflections. The MIT Press, Cambridge, pp 227–246
Walsh D, Lewens T, Ariew A (2002) The trials of life natural selection and random drift. Philos Sci 69(3):429–446
West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, Oxford, New York