



HELLENIC REPUBLIC

**National And Kapodistrian
University Of Athens**

SCHOOL OF SCIENCE

DEPARTMENT OF HISTORY AND PHILOSOPHY OF SCIENCE

DOCTORAL DISSERTATION

**Levels, Decomposition and Mechanistic Explanation in
Biological Systems: The Case of Cognitive Science**

Amalia K Tsakiri

ATHENS

SEPTEMBER 2024



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

**Εθνικό και
Καποδιστριακό
Πανεπιστήμιο Αθηνών**

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΙΣΤΟΡΙΑΣ ΚΑΙ ΦΙΛΟΣΟΦΙΑΣ ΤΗΣ ΕΠΙΣΤΗΜΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Επίπεδα, Αποσύνθεση και Μηχανιστική Εξήγηση στα Βιολογικά

Συστήματα: Η Περίπτωση της Γνωσιακής Επιστήμης

Αμαλία Κ Τσακίρη

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2024

PhD DISSERTATION

Levels, Decomposition and Mechanistic Explanation in Biological Systems: The Case of Cognitive Science

Amalia K Tsakiri

SUPERVISOR: Drakoulis Nikolinakos, Professor, University of Athens

THREE-MEMBER ADVISORY COMMITTEE:

Drakoulis Nikolinakos, Professor, University of Athens

Stathis Psillos, Professor, University of Athens

Eleni Manolakaki, Assistant Professor, University of Athens

SEVEN-MEMBER EXAMINATION COMMITTEE

Drakoulis Nikolinakos
Professor, UoA

Stathis Psillos
Professor, UoA

Eleni Manolakaki
Assistant Professor, UoA

Irini Skaliora
Professor, UoA

Anthony Hatzimoysis
Professor, UoA

Stavros Ioannidis
Assistant Professor, UoA

Orly Shenker
Professor,
the Hebrew University of Jerusalem

Examination Date: 25/10/2024

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Επίπεδα, Αποσύνθεση και Μηχανιστική Εξήγηση σε Βιολογικά Συστήματα:

Η Περίπτωση της Γνωσιακής Επιστήμης

Αμαλία Κ Τσακίρη

ΕΠΙΒΛΕΠΩΝ: Δρακούλης Νικολινάκος, Καθηγητής, ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ:

Δρακούλης Νικολινάκος, Καθηγητής, ΕΚΠΑ

Στάθης Ψύλλος, Καθηγητής, ΕΚΠΑ

Ελένη Μανωλακάκη, Επίκουρη Καθηγήτρια, ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Δρακούλης Νικολινάκος
Καθηγητής, ΕΚΠΑ

Στάθης Ψύλλος
Καθηγητής, ΕΚΠΑ

Ελένη Μανωλακάκη
Επίκουρη Καθηγήτρια, ΕΚΠΑ

Ειρήνη Σκαλιόρα
Καθηγήτρια, ΕΚΠΑ

Αντώνης Χατζημωϋσής
Καθηγητής, ΕΚΠΑ

Σταύρος Ιωαννίδης
Επίκουρος Καθηγητής, ΕΚΠΑ

Orly Shenker
Καθηγήτρια,
the Hebrew University of Jerusalem

Ημερομηνία Υποστήριξης: 25/10/2024

License Information

© 2025 Amalia Tsakiri

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

You are free to share, copy, and redistribute the material in any medium or format under the following terms:

- Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
- NonCommercial — You may not use the material for commercial purposes.
- NoDerivatives — If you remix, transform, or build upon the material, you may not distribute the modified material.
- No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

To view a copy of this license, visit: <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Πληροφορίες Άδειας Χρήσης

© 2025 Αμαλία Τσακίρη

Το παρόν έργο διατίθεται υπό την άδεια Creative Commons Αναφορά-Μη Εμπορική Χρήση-Χωρίς Παράγωγα Έργα 4.0 Διεθνές (CC BY-NC-ND 4.0).

Επιτρέπεται η αναπαραγωγή, η αποθήκευση και η διανομή του έργου, υπό τις ακόλουθες προϋποθέσεις:

- Αναφορά — Πρέπει να αποδώσετε την πρέπουσα αναγνώριση στον δημιουργό, να παρέχετε σύνδεσμο προς την άδεια και να επισημάνετε αν έγιναν αλλαγές.
- Μη Εμπορική Χρήση — Δεν μπορείτε να χρησιμοποιήσετε το έργο για εμπορικούς σκοπούς.
- Χωρίς Παράγωγα Έργα — Αν τροποποιήσετε, επεξεργαστείτε ή βασιστείτε στο έργο, δεν επιτρέπεται να διανείμετε το τροποποιημένο υλικό.
- Χωρίς πρόσθετους περιορισμούς — Δεν μπορείτε να εφαρμόσετε νομικούς όρους ή τεχνολογικά μέτρα που να περιορίζουν νομικά τρίτους από το να κάνουν οτιδήποτε επιτρέπει η άδεια.

Για να δείτε ένα αντίγραφο της άδειας: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.el>

ABSTRACT

Few phenomena are as intriguing and enigmatic as the ones that constitute mentality, and investigating them as behaviors of biological organisms (humans, in this case) poses an enormous challenge for the sciences of the mind. These sciences, collectively known as cognitive science, form an umbrella that embraces the study of mental phenomena. It is in this field that the idea for the new-mechanistic model of explanation tests its limits, as it must not only tame the explanatory complexity of biology but also delve deep into the mysteries of the mind. It is into these deep intellectual waters that the present dissertation descends.

This inquiry targets various interconnected issues: Can the new-mechanistic explanatory framework be extended to explain mental phenomena? How can we reconcile the diverse concepts of *levels* encountered in cognitive science with mechanistic levels? What role does *decomposition* play in explaining biological and mental phenomena, primarily as an analytical tool and also as an integral part of the mechanistic explanatory schema? Can mechanistic explanatory models incorporate the informational and representational aspects of cognitive processes? Ultimately, can the gap between psychological and neurobiological approaches to cognition be bridged with the aid of the mechanistic explanatory framework?

The research unfolds across various stages, progressing from the general to the specific. This gradual specialization is reflected in the two parts of the study, the sequence of chapters, and their internal structure. The research methodologies employed to address these questions approach the subject matter from multiple angles. The study begins with conceptual mapping and analysis of the two fundamental components that constitute the research axis of the dissertation. These are the concepts of levels and decomposition, which are examined in the first two chapters. The subsequent two chapters present a comparative analysis of two explanatory models: Carl Gillett's (2016) *compositional explanations* and Carl Craver's (2007) *new-mechanistic explanatory framework*, which focuses on neuroscience, with an emphasis on constitutive explanations. These models are closely related, as they both emphasize the role of recursive decomposition leading to a hierarchy of multiple explanatory levels. These two chapters are framed by two extensive case studies. The first, presented in the third chapter, concerns compositional explanations of the bending mechanism of the knee and the physiology of muscle contraction. The second, detailed in the fourth chapter, explores the constitutive mechanistic explanation of neurobiological mechanisms of spatial memory development in mice.

While the first part lays the groundwork for mechanistic explanation by initially examining levels, decomposition, and biological complexity, the second part delves into the deep waters of applying mechanistic explanations to mental phenomena. It begins by analyzing their distinctive character, which stems from information processing and intentionality -the capacity for internal representation of the world and the attribution of meaning to it. A comparative study is conducted to investigate the compatibility of levels used in cognitive science explanations with mechanistic explanatory levels, using visual recognition of an apple as a case study. The fifth chapter concludes with a negative assessment regarding levels, constituting the first major challenge to the applicability of mechanistic explanation to mental phenomena.

The final chapter presents a comparative and critical discussion on the possibility of unifying psychological and mechanistic models into a common explanatory framework, highlighting the problems that arise in such an endeavor. The concept of decomposition is then revisited as an analytical tool for constructing psychological explanatory models, examined in the case of the internal visual representation of a horse. Finally, it highlights the issues that emerge when applying this to physical (neurobiological) and mental (concerning informational and representational content) entities. The chapter and the dissertation conclude with a proposal for reconciliation between psychological and neurobiological explanatory models. In this proposal, psychological models play a central role, with mechanistic explanations subsumed within them as a hierarchy of explanatory mechanisms extending across many lower levels. Each hierarchy of mechanisms implements a specific explanatory level of the psychological model, without a one-to-one correspondence between neurological and psychological elements that recursively extends to all lower levels.

Concluding the study, while mechanistic explanations offer valuable insights for understanding mental phenomena, they require significant modifications to fully capture their unique characteristics. The proposed framework suggests a way to integrate psychological and neurobiological approaches to cognition without reductionism, adopting a form of interpretive pluralism that respects the autonomy of different levels and aspects of explanation while illuminating their interconnections.

SUBJECT AREA: Philosophy of Mind and Cognitive Science, Philosophy of Biology

KEYWORDS: new-mechanism, near-decomposability, integrated explanations, biological complexity, psychological models, information processing

ΠΕΡΙΛΗΨΗ

Λίγα φαινόμενα είναι τόσο ενδιαφέροντα και τόσο δισεπίλυτα όσο αυτό της νοητικότητας και η διερεύνησή του ως συμπεριφορά βιολογικών οργανισμών (ανθρώπων, εν προκειμένω) αποτελεί μια τεράστια πρόκληση για τις επιστήμες του νου, που ως μια ομπρέλα που αγκαλιάζει την έρευνα των νοητικών φαινομένων συνιστούν την γνωσιακή επιστήμη. Σε αυτό το πεδίο είναι που η πρόταση για το νέο-μηχανιστικό μοντέλο εξήγησης δοκιμάζει τα όριά της, καθώς δεν έχει μόνο να δαμάσει εξηγητικά την βιολογική πολυπλοκότητα, αλλά και να κατανοήσει σε βάθος τα μυστικά του νου. Σε αυτά τα βαθιά ερευνητικά νερά καταδύεται η παρούσα διατριβή.

Η μελέτη βάζει στο στόχαστρο διάφορα αλληλένδετα ζητήματα: Μπορεί το νέο-μηχανιστικό εξηγητικό πλαίσιο να επεκταθεί στην εξήγηση των νοητικών φαινομένων; Πώς μπορούμε να συμφιλιώσουμε τις διάφορες έννοιες των *επιπέδων* που συναντούμε στη γνωσιακή επιστήμη με τα μηχανιστικά επίπεδα; Ποιος είναι ο ρόλος της *αποσύνθεσης* στην εξήγηση βιολογικών και νοητικών φαινομένων, πρώτα από όλα ως αναλυτικό εργαλείο και επίσης ως αναπόσπαστο τμήμα του μηχανιστικού εξηγητικού σχήματος; Μπορούν τα μηχανιστικά εξηγητικά μοντέλα να ενσωματώσουν τις πληροφοριακές και αναπαραστασιακές πτυχές των νοητικών διεργασιών; Μπορεί, εν τέλει, να γεφυρωθεί το χάσμα μεταξύ ψυχολογικών και νευροβιολογικών προσεγγίσεων της νόησης με την βοήθεια του μηχανιστικού πλαισίου εξήγησης;

Η έρευνα ξετυλίγεται σε διάφορα στάδια ξεκινώντας από το γενικότερο και καταλήγοντας στο ειδικότερο. Αυτή η σταδιακή εξειδίκευση αντανακλάται στα δύο μέρη της έρευνας, στην σειρά των κεφαλαίων και στην εσωτερική τους διάρθρωση. Οι ερευνητικές μεθοδολογίες που επιστρατεύτηκαν για να απαντηθούν όλα τα προαναφερθέντα ερωτήματα προσέγγισαν το ζητούμενο από διάφορες πλευρές. Η έρευνα ξεκινά με εννοιολογική χαρτογράφηση και ανάλυση των δύο βασικών συνιστωσών που συγκροτούν τον ερευνητικό άξονα της διατριβής. Αυτές είναι η έννοια των επιπέδων και της αποσύνθεσης και μελετώνται στα δύο πρώτα κεφάλαια. Στα επόμενα δύο κεφάλαια ακολουθεί η συγκριτική παρουσίαση δύο εξηγητικών μοντέλων: των συστατικών εξηγήσεων (compositional explanations) του Carl Gillett (2016) και του νέο-μηχανιστικού εξηγητικού πλαισίου με έμφαση στις συγκροτητικές (constitutive) εξηγήσεις του Carl Craver (2007), που εστιάζει στις νευροεπιστήμες. Τα μοντέλα συνδέονται στενά, γιατί δίνουν σημασία στον ρόλο της αναδρομικής (recursive) αποσύνθεσης που οδηγεί σε μια ιεραρχία πολλών εξηγητικών επιπέδων. Τα δύο αυτά κεφάλαια πλαισιώνονται με δύο εκτεταμένες μελέτες περίπτωσης που αφορούν συστατικές εξηγήσεις του καμπτικού μηχανισμού του αγκώνα και της φυσιολογίας της μυϊκής σύστολης (στο τρίτο κεφάλαιο) και στην παρουσίαση της συγκροτητικής μηχανιστικής εξήγησης στους νευροβιολογικούς μηχανισμούς ανάπτυξης χωρικής μνήμης σε ποντίκια (στο τέταρτο).

Ενώ το πρώτο μέρος στρώνει το έδαφος για τη μηχανιστική εξήγηση, εξετάζοντας αρχικά τα επίπεδα, την αποσύνθεση και τη βιολογική πολυπλοκότητα, το δεύτερο μέρος μπαίνει στα βαθιά νερά της εφαρμογής των μηχανιστικών εξηγήσεων στα νοητικά φαινόμενα. Αρχικά, αναλύεται ο ιδιαίτερος χαρακτήρας τους που οφείλεται στην επεξεργασία πληροφορίας και την προθετικότητα, δηλαδή την ικανότητα εσωτερικής αναπαράστασης του κόσμου και την νοηματοδότηση της αλληλεπίδρασης με αυτόν. Εφαρμόζεται συγκριτική μελέτη ώστε να διερευνηθεί η συμβατότητα των επιπέδων που χρησιμοποιούνται στις εξηγήσεις εντός της γνωσιακής επιστήμης με τα μηχανιστικά εξηγητικά επίπεδα και παρουσιάζεται ως μελέτη περίπτωσης η οπτική αναγνώριση ενός μήλου. Το πέμπτο κεφάλαιο καταλήγει σε μια αρνητική αποτίμηση σε σχέση με τα επίπεδα, που συνιστά και την πρώτη μεγάλη πρόκληση της εφαρμοσιμότητας της μηχανιστικής εξήγησης στα νοητικά φαινόμενα.

Στο τελευταίο κεφάλαιο παρουσιάζεται συγκριτικά και κριτικά η συζήτηση περί δυνατότητας ενοποίησης ψυχολογικών και μηχανιστικών μοντέλων σε ένα κοινό εξηγητικό μοντέλο και αναδεικνύονται τα προβλήματα που προκύπτουν σε μια τέτοια περίπτωση. Στην συνέχεια επανέρχεται στην συζήτηση η έννοια της αποσύνθεσης ως αναλυτικό εργαλείο για την κατασκευή ψυχολογικών εξηγητικών μοντέλων, μελετάται στην περίπτωση της εσωτερικής οπτικής αναπαράστασης ενός αλόγου και τέλος, αναδεικνύονται τα προβλήματα που προκύπτουν από την εφαρμογή της σε φυσικές (νευροβιολογικές) και νοητικές (που αφορούν σε πληροφοριακό και αναπαραστασιακό περιεχόμενο) οντότητες. Το κεφάλαιο, και η διατριβή, καταλήγει με μια πρόταση συμφιλίωσης μεταξύ των ψυχολογικών και νευροβιολογικών εξηγητικών μοντέλων όπου κεντρικό ρόλο παίζουν τα ψυχολογικά μοντέλα και οι μηχανιστικές εξηγήσεις υπάγονται σε αυτά ως μια ιεραρχία εξηγητικών μηχανισμών που εκτείνονται σε πολλά, χαμηλότερα, επίπεδα και κάθε ιεραρχία μηχανισμών υλοποιεί ένα συγκεκριμένο επίπεδο εξήγησης του ψυχολογικού μοντέλου, χωρίς να υπάρχει μια 1:1 αντιστοίχιση μεταξύ νευροβιολογικών και ψυχολογικών στοιχείων που να εκτείνεται, αναδρομικά εφαρμοζόμενη, σε όλα τα κατώτερα επίπεδα.

Εν κατακλείδι, ενώ οι μηχανιστικές εξηγήσεις προσφέρουν πολύτιμες γνώσεις για την κατανόηση των νοητικών φαινομένων, χρειάζονται σημαντικές τροποποιήσεις για να συλλάβουν πλήρως τα μοναδικά χαρακτηριστικά τους. Το προτεινόμενο πλαίσιο υποδεικνύει έναν τρόπο ενσωμάτωσης ψυχολογικών και νευροβιολογικών προσεγγίσεων της νόησης χωρίς αναγωγισμό, υιοθετώντας μια μορφή ερμηνευτικού πλουραλισμού που σέβεται την αυτονομία διαφορετικών επιπέδων και πτυχών εξήγησης, ενώ παράλληλα διαφωτίζει τις διασυνδέσεις τους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Φιλοσοφία του Νου και της Γνωσιακής Επιστήμης και Φιλοσοφία της Βιολογίας

ΛΕΞΕΙΣ - ΚΛΕΙΔΙΑ: μηχανισμός, αποσυνθεσιμότητα, εξηγητική ενοποίηση, βιολογική πολυπλοκότητα, ψυχολογικά μοντέλα, επεξεργασία πληροφορίας

To my beloved husband and my precious little twin-devils

TABLE OF CONTENTS

INTRODUCTION	1
Background and Context	2
Research Problem.....	3
Significance of the Study	4
Theoretical Framework	5
Research Questions and Objectives	6
Methodology Overview	7
Structure of the Dissertation	9
Scope and Limitations	14
Conclusion and Acknowledgments.....	15
PART I FOUNDATIONS OF MECHANISTIC EXPLANATION: LEVELS, DECOMPOSITION, AND BIOLOGICAL COMPLEXITY.....	18
CHAPTER ONE - HIERARCHY AND LEVELS IN BIOLOGICAL ORGANIZATION.....	19
1.0. Outline	19
1.1. Laying the Groundwork: Introducing Hierarchy and Levels	19
1.2. Levels in Living Systems under a Biological Perspective	23
1.2.1. Levels in science and nature.....	24
1.2.2. Levels in neurobiology.....	25
1.2.3. Extended levels.....	27
1.2.4. Levels that deal with complexity	30
1.3. Summary and conclusions.....	34
CHAPTER TWO - DECOMPOSITION AS A TOOL FOR ANALYZING AND UNDERSTANDING BIOLOGICAL SYSTEMS	36
2.0. Outline	36
2.1. From Parts to Whole and Back Again: Exploring Composition and Decomposition	37
2.2. Breaking Down the Complex: Decomposition in Biological Systems	41
2.2.1. Understanding Complexity in Biological Systems	41
2.2.2. Approaches to Decomposing Biological Systems	45
2.3. Structural Principles and Organizational Dynamics in Complex Biological Systems	49
2.3.1. Aggregativity vs Emergence: When Parts Become More Than Their Sum.....	51
2.3.2. Interactivity and Modular Organization in Biological Structures	58
2.4. Idealization and Near-Decomposability: Bridging Simplicity and Complexity	63

2.5. Summary and Conclusions	74
CHAPTER THREE - EXPLORING BIOLOGICAL MECHANISMS VIA COMPOSITIONAL EXPLANATIONS 76	
3.0. Outline	76
3.1. The Evolution of Explanatory Models in Biological Sciences.....	76
3.1.1. From Laws to Mechanisms: Shifting Paradigms in Biology	78
3.1.2. Statistical Approaches and Their Limitations in Biological Understanding	82
3.1.3. Decomposition and Multilevel Analysis as Basis for Compositional Explanations.....	85
3.2. Gillett's Framework on <i>Scientific Composition</i>	88
3.3. Elbow Movement as a Model for Multilevel Biological Analysis.....	94
3.3.1. Anatomical and Physiological Essentials of Forearm Flexion	95
3.3.2. Decomposing the Elbow Movement System to Reveal Hierarchical Organization.....	98
3.3.3. Muscle Contraction Mechanisms Through the Lens of Compositional Explanation.....	103
3.4. Summary and Conclusions	110
CHAPTER FOUR - DECOMPOSING PHENOMENA INTO MECHANISMS IN THE NEW-MECHANISTIC APPROACH	
4.0. Outline	112
4.1. Tracing the Roots of New-Mechanism	112
4.1.1. Conceptualizing New-Mechanism	113
4.1.2. The Evolution of New-Mechanistic Thought	116
4.1.3. Craver's Account of Constitutive Mechanistic Explanation in Neuroscience	124
4.2. What is and what is not a mechanism?.....	128
4.2.1. Distinguishing Mechanisms: Various Interpretations	129
4.2.2. Boundaries and Challenges of Mechanistic Explanation.....	133
4.3. Mechanistic Decomposition: A Case Study in Spatial Memory.....	139
4.3.1. Framing the Explanandum: Defining Spatial Memory	141
4.3.2. Levels of Analysis in Spatial Navigation	144
4.3.3. From Behavior to Brain: Decomposing the Hippocampal Mechanism	145
4.3.4. Cellular and Subcellular Mechanisms of Memory Formation	147
4.4. Summary and Conclusions	149
PART II: EXPLORING THE APPLICABILITY OF MECHANISTIC EXPLANATIONS TO MENTAL PHENOMENA	
CHAPTER FIVE – MULTIPLE LEVELS AND COGNITIVE SCIENCE EXPLANATIONS 153	
5.0. Outline	153

5.1. The Unique Nature of Mental Phenomena in Explanatory Frameworks	154
5.1.1. Intentionality and Information Processing: Defining Features of Mental Phenomena	155
5.1.2. Information and Representation in Cognitive Systems.....	160
5.1.3. The Vehicle-Content Distinction as a Key-aspect of Cognitive Explanations	165
5.2. Navigating the Levels: The Role of Multiple Perspectives in Cognitive Science.....	167
5.2.1. An Overview of the Diverse Notions of Levels in Mental Phenomena Research.....	168
5.2.2. Visual Perception as a Model for Multilevel Explanations in Cognitive Science	182
5.3. Reconciling Cognitive Science Explanations with Mechanistic Levels.....	189
5.3.1. Revisiting Levels of Mechanisms: Strengths and Limitations	190
5.3.2. The Challenge of Incorporating Informational Content in Levels of Mechanisms	194
5.4. Summary And Conclusions	199
CHAPTER SIX - COGNITIVE MECHANISMS: SO CLOSE, YET SO FAR	202
6.0. Outline	202
6.1. The Debatable Viability of Mechanistic Explanation in the Cognitive Realm	203
6.1.1 Against the Autonomy of Psychological Explanations.....	206
6.1.2. Complementarity, but not Seamless Integration	211
6.1.3. Employing and Expanding Marr's Trilevel Account	216
6.2. Is There Any Place for Mental Properties in Mechanistic Hierarchies?	220
6.3. Recursive Decomposition as a Key to Multilevel Cognitive Explanations	225
6.3.1. Applying Recursive Decomposition to Information Processing Systems	227
6.3.2. A Case Study in Visual Perception Demonstrating Recursive Informational Decomposition	232
6.4. Towards “Cognitive Rich” Mechanistic Hierarchies in Explanatory Models	241
6.5. Summary And Conclusions	252
CONCLUSIONS	255
Summary.....	255
Conclusions by Chapter	256
General Conclusions	262
Negative account.....	262
Positive account.....	263
REFERENCES.....	265

LIST OF FIGURES

- Figure 1* (p. 57): Formation of a hierarchical structure through sub-aggregation.
- Figure 2* (p. 58): 'Super-agents': Grouping of individuals into single objects that represent the collective.
- Figure 3* (p. 66): The idealized case of a completely decomposable system.
- Figure 4* (p. 69): A nearly-decomposable system illustrating heat diffusion in a thermally insulated building.
- Figure 5* (p. 71): Difference between a completely decomposable system and a near-decomposable system.
- Figure 6* (p. 104): Organization of muscle.
- Figure 7* (p. 108): The sliding filament model.
- Figure 8* (p. 109): The "walk-along" mechanism for contraction of the muscle.
- Figure 9* (p. 140): A phenomenon (top) and its mechanism (bottom).
- Figure 10* (p. 141): Three levels of mechanisms.
- Figure 11* (p. 141): Levels are defined locally within decomposition hierarchies.
- Figure 12* (p. 146): The entorhinal/hippocampal system.
- Figure 13* (p. 149): Four mechanistic levels for the explanation of the behavior of a mouse, navigating a Morris water maze.
- Figure 14* (p. 231): Recursive decomposition of mental processes.
- Figure 15* (p. 233): Representation of entire objects in high-level visual processing.
- Figure 16* (p. 235): A trilevel decomposition of visual perception, with an emphasis on high-level processing.

INTRODUCTION

Few scientific fields could be more puzzling than those dealing with the core of human existence - our ability to think, perceive the world around us, hold beliefs, make plans for the future, and reminisce about past events. In short, our capacity for mentality. We can do all these things using one of evolution's most complex constructions: the brain. It is an outstanding example of biological organization and studying it sometimes feels like diving into the depths of an unknown ocean. In our attempt to explore the secrets of the mind, we must simultaneously tame the complexity of the human brain and correlate it with the mental phenomena that emerge from its functioning.

In recent years, it has become increasingly apparent that intelligence is not necessarily tied exclusively to biological organisms or to the kind of brains possessed by animals that exhibit relatively advanced cognitive functions, an insight that has been reinforced by rapid developments in artificial intelligence and research on organisms like octopuses, which exhibit strong signs of higher intelligence without following the basic brain-centric structure found in humans and most representatives of the animal kingdom. However, for the purposes of this dissertation, I will focus on the intelligence manifested in biological organisms and will explore how mental phenomena can be explained in relation to our understanding of the complexity and multi-leveled nature of biological systems.

The present inquiry is part of a research venture that spans the fields of philosophy of biology, complexity, mind, and cognitive neuroscience, with the common denominator being the need to find an explanatory framework that can be applied to biologically instantiated cognitive phenomena. In recent years, the necessity for a different explanatory model for biological and neurobiological phenomena has been highlighted by the New-Mechanistic movement in the philosophy of science. Motivated by the very particular nature of biological systems and their behaviors, this movement has raised strong doubts about whether traditional explanatory schemas applied in other sciences such as physics, chemistry, and the like, can be applied to biological phenomena. Consequently, new-mechanists proposed an entirely different explanatory approach. This approach aims to provide a framework for understanding how sciences dealing with biological and neurobiological phenomena offer explanations and to serve as a tool for evaluating these explanations.

Background and Context

Cognitive science, an interdisciplinary field that emerged in the mid-20th century, seeks to understand the nature of the mind and its processes. It brings together insights from psychology, neuroscience, philosophy, linguistics, anthropology, and computer science to explore how information is represented, processed, and transformed in cognitive systems. Central to this endeavor is the challenge of explaining mental phenomena - the subjective experiences, thought processes, and behaviors that characterize human cognition.

Understanding mental phenomena is a complex task that requires multiple perspectives and explanatory frameworks. Traditional approaches in psychology have often relied on functional analyses and information processing models, while neuroscience has focused on uncovering the biological mechanisms underlying cognitive functions. Philosophy of mind has grappled with questions of consciousness, intentionality, and the relationship between mind and brain. More recently, the rise of computational approaches has introduced new ways of conceptualizing and modeling cognitive processes. The diversity of these approaches reflects the multifaceted nature of cognition itself. Mental phenomena span multiple levels of organization, from molecular interactions in neural circuits to high-level cognitive functions like decision-making and language comprehension. This multilevel character of cognition presents a unique challenge: How can we develop explanatory frameworks that integrate insights from these various levels and disciplines?

In recent decades, the new-mechanistic approach to explanation has gained prominence in the philosophy of science, particularly in the biological sciences. This approach, which focuses on identifying and describing the mechanisms underlying phenomena, has shown promise in providing integrated explanations across multiple levels of organization. However, its application to cognitive phenomena raises important questions about the nature of mental processes and how psychological and neurobiological explanations could be related.

In this dissertation, I explore the potential and challenges of applying mechanistic explanations to cognitive phenomena, with a particular focus on the concepts of levels and decomposition. By examining how these concepts are understood and applied in both cognitive science and mechanistic philosophy, I aim to contribute to the ongoing dialogue about how best to explain the complexities of the mind.

Research Problem

The central issue I address in this dissertation is the challenge of applying constitutive mechanistic explanations to cognitive phenomena. While the mechanistic approach has proven successful in explaining many biological processes, its application to mental phenomena presents unique difficulties. These challenges stem from the distinctive nature of cognitive systems, which process information and exhibit intentionality, setting them apart from other biological systems.

The present inquiry deals with several interconnected problems. Cognitive phenomena span multiple levels of organization, from neural to brain processes, and at the same time they are analyzed and described on multiple levels. Mechanistic explanations must account for this multilevel nature, but it's unclear how to integrate all the diverse notions of levels within a single explanatory framework. Reconciling these different conceptions of levels with the mechanistic notion of levels presents a significant challenge, too. Furthermore, unlike purely physical systems, cognitive systems process information and generate internal representations. Mechanistic explanations need to accommodate these informational and representational aspects, which are not easily reducible to physical mechanisms.

In philosophy of mind, a crucial distinction is between the vehicles of mental content (e.g., neural states or forms of representations) and the content itself (e.g., beliefs, perceptions). Mechanistic explanations must navigate this distinction, explaining both how physical mechanisms implement cognitive processes and how they give rise to meaningful mental content. There's an ongoing debate about whether psychological explanations can be autonomous from neurobiological explanations. The application of mechanistic explanations to cognitive phenomena must address this issue, determining whether and how psychological and neurobiological explanations can be integrated.

Another important issue is that mechanistic explanations often rely on the decomposition of systems into component parts and operations. However, the extent to which cognitive systems can be decomposed without losing essential properties is a matter of debate. This issue is closely linked to the complexity exhibited by biological systems in general and cognitive systems in particular. Being systems that display increased interactivity among their component parts and the ability to form structural or dynamical functional modules, they resist explanatory strategies that assume their complete decomposability without at the same time excluding that they can be decomposed. There is an ongoing discussion about whether the idea of treating these systems as nearly

decomposable can, to some degree, circumvent this problem, allowing for their study as decomposable systems, even if only approximately. It is with this reasoning that they are approached in the present investigation as well.

This dissertation aims to address these challenges, exploring whether and how mechanistic explanations can be adapted or extended to provide comprehensive accounts of cognitive phenomena that respect both their physical implementation and their distinctive mental characteristics.

Significance of the Study

My investigation delves into both the philosophy of science and cognitive science. In terms of the philosophy of science, it touches on crucial issues concerning the scope of applicability of the mechanistic model of explanation across various scientific inquiries. I aim to evaluate whether mental phenomena represent the ultimate boundary of this explanatory model or if it can potentially contribute to ongoing efforts to bridge the gap between psychological and neurobiological approaches to understanding the mind. By examining how explanatory frameworks, primarily developed for biological systems, might be adapted or extended to account for the unique characteristics of cognitive phenomena, my exploration offers important insights into the nature of scientific explanation itself. It also sheds light on the limitations of mechanistic approaches across different domains of inquiry and raises thought-provoking questions about the nature of explanation in science and the extent to which we can apply uniform explanatory models across diverse fields of study.

Regarding cognitive science, the significance of my work lies in its potential to advance our understanding of cognitive phenomena in several ways. First, by critically examining the challenges of applying mechanistic explanations to mental phenomena, my research helps to clarify the unique characteristics of cognitive systems that set them apart from other biological systems in relation to their compatibility with these types of explanations. This clarification is crucial for developing explanatory frameworks that can adequately capture the uniqueness and complexity of mental phenomena. Second, my inquiry contributes to exploring and clarifying the differences and relations between various notions of levels used in cognitive science explanations. By exploring how mechanistic explanations can potentially integrate multiple levels of analysis - from neural mechanisms to high-level cognitive functions - we gain insights into how we might develop more comprehensive, multilevel accounts of cognition. Third, there are also practical implications for

cognitive science methodology. The critical evaluation of the concept of decomposition in the context of cognitive systems can inform research strategies and experimental designs in cognitive neuroscience and psychology, potentially leading to more effective approaches for studying mental phenomena. Lastly, by addressing the tension between the autonomy of psychological explanations and the integration of psychological and neurobiological accounts, my research contributes to ongoing debates about the relationship between different scientific disciplines and the possibility of developing unified theories in cognitive science.

Theoretical Framework

In this dissertation, I draw upon and integrate several theoretical frameworks and concepts from cognitive science, philosophy of science, and the study of complex systems. At its core, my inquiry engages with the new-mechanistic philosophy, a contemporary approach to scientific explanation that focuses on identifying and describing the mechanisms underlying phenomena. This framework, developed by philosophers such as Machamer, Darden, Craver, Glennan, and Bechtel among others, provides a foundation for understanding how complex systems produce observable behaviors through the organized activities of their component parts. In the new-mechanistic framework, a widely accepted definition of a mechanism is: “A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon (Glennan & Illari, 2017a, p. 2).

In the development of the mechanistic explanatory framework, two distinct approaches have emerged: the etiological and the constitutive. The former focuses on how a mechanism brings about or causes a phenomenon, while the latter delves into how it constitutes it. I am primarily addressing the constitutive approach, aiming to elucidate the process of decomposing a system into its constituent parts. To shed light on this concept, I introduce Gillett's account of scientific composition. This framework helps us understand how components at a lower level, each with a specific functional role, collectively give rise to the phenomenon being explained. Gillett posits that compositional explanations are “explanations of higher-level entities built around lower-level entities taken to compose them” (Gillett, 2016, p. 2).

Central to this study and related to the former emphasis on explanations that extend to many compositional levels is the effort of conceptualizing levels of explanation, which is examined from multiple perspectives. I consider various notions of levels encountered in cognitive science, including levels of organization (from molecules to brain regions and whole organisms), levels of

processing (from sensory inputs to high-level cognition), and levels of analysis (such as Marr's computational, algorithmic, and implementational levels). These are contrasted and compared with the concept of mechanistic levels, as articulated by Craver (2007), which focuses on local part-whole relations within mechanisms.

The Information Processing Approach to Cognition (Palmer & Kimchi, 1986) serves as another theoretical contribution. This framework, rooted in the cognitive revolution of the mid-20th century, conceptualizes mental processes as computations performed on representations. I examine how this approach relates to mechanistic explanations, particularly concerning the way that information processing can be understood in terms of physical mechanisms. The closely related idea of functional analysis, as developed by Cummins (1975), and functional decomposition, provide a framework for understanding how complex capacities can be explained in terms of the organized capacities of a system's parts. This is considered in relation to mechanistic explanations and the challenges of decomposing cognitive systems.

A large part of the dissertation is devoted to the theoretical investigation of complexity as it manifests in biological systems, with a focus on the role of decomposition as a tool for studying them. This part touches on the core of biological systems which, due to their inherent nature, simultaneously exhibit properties that make them decomposable and properties that contradict their decomposition. This occurs in systems where there is a high degree of interactivity between their constituent parts, which, however, presents significant gradations, resulting in the creation of 'islands of order within chaos', to use a more metaphorical language. In this work, I could not avoid addressing this 'chaotic' topic, given that the two main pillars of mechanistic explanation through which the research is interwoven are decomposition and levels, and for understanding the first this particular investigation was inevitable.

Research Questions and Objectives

In this research, I aim to critically examine the applicability of mechanistic explanations to cognitive phenomena, with the ultimate goal of determining whether and how such explanations can provide a comprehensive account of mental processes that respect both their physical implementation and their distinctive mental characteristics. To this end, my research is guided by several interconnected questions and objectives.

A central focus is to explore whether the new-mechanistic framework, primarily developed for biological sciences, can be effectively extended to explain cognitive phenomena, and if so, what

modifications or extensions might be necessary to account for the unique features of mental processes. This is closely tied to the question of how the various notions of 'levels' employed within cognitive science can be reconciled with the concept of mechanistic levels, seeking to explore whether a unified understanding of levels can be developed that accommodates both the hierarchical organization of physical mechanisms and the levels often used in cognitive explanations.

An important objective is to address how mechanistic explanations can incorporate the informational and representational aspects of cognitive processes, tackling the challenge of integrating the physical and semantic dimensions of cognition within a single explanatory framework. Could mechanistic explanations bridge the gap between psychological and neurobiological accounts of cognition?

Through addressing these questions and objectives, I aim to construct a philosophical argument about the nature of explanation in cognitive science and seek to determine whether a modified or extended version of mechanistic explanation can provide an integrated framework for understanding cognitive phenomena across multiple levels of organization and analysis.

Methodology Overview

To achieve the objectives of this inquiry I used several research methodologies. I combined conceptual analysis with critically examining and comparing theories and models, and I thoroughly explored the application of these insights to specific case studies in kinesiology, neuroscience, and cognitive neuroscience. Lastly, I proposed a solution to the problems that arise when attempting to integrate mechanistic with psychological models, by developing a framework under which their unification could be achieved.

More specifically, I first focused on the conceptual mapping of the two central axes of the dissertation: the concept of levels and decomposition. I initially examined how levels of organization and hierarchies of levels are understood within the biological field, using the conceptual framework of systems theory and mereology. To better understand the concept of levels, I presented and compared the main accounts on categorizing biological levels of organization based on several criteria. This conceptual analysis was complemented in the second part of the dissertation by focusing on the levels used in methodological approaches and explanations carried out within cognitive science. The second concept I examined is that of decomposition. In the first part of the dissertation, I explored how it is understood as an analytical

tool in biological sciences and how it relates to the concept of complexity. In the second part, which focuses on cognitive science, I examined it within the framework of the Information Processing Approach to Cognition, which uses recursive decomposition of informational events to construct psychological models for explanation. Finally, I comparatively examined the basic concepts that constitute the central core of mentality, namely the distinctions between information and representation, and between content and vehicle of a mental state. The above conceptual analysis was necessary to understand why the rationale of mechanistic explanation is compatible with the study of biological systems and how it could also be compatible with the phenomena studied by cognitive science, as well as what the difficulties are regarding its application to the latter.

A second axis around which my methodology revolved is the comparative analysis of explanatory models used in biological and cognitive phenomena. I examined in depth the framework of compositional explanations proposed by Gillett (2016), to proceed, next to the mechanistic model of explanation as proposed mainly by Craver (2007) and Bechtel (2008). In this way, I demonstrated how the constitutive explanations proposed by new-mechanists fit into the broader framework of compositional explanations. I also referred to functional analysis as proposed by Cummins (1975; 1983), Marr's trilevel analysis of information processing systems (1982), and finally, I extensively examined the explanatory approach of the Information Processing Approach to Cognition as presented by Palmer and Kimchi (1986), the latter aiming to highlight the role of recursive functional decomposition in the analysis of mental events.

A third crucial component of my methodology involves case studies from the fields of biological and cognitive sciences. Specifically, I used a kinesiological example (the elbow flexion movement) to demonstrate how Gillett's compositional explanations can be understood in a multi-level hierarchy, starting from the highest embodied level and culminating in the detection of molecular mechanisms underlying muscle contraction. To illustrate how the new-mechanistic account performs constitutive explanations across a hierarchy of levels, I employed the example of spatial memory in mice. Here, I attempted to combine the use of decomposition of the phenomenon and system into entities and activities while simultaneously highlighting how these connect with the mechanistic levels framework proposed by Craver (2007). Moving, now, to the section specializing in cognitive science, I focused on various mechanisms within the visual system that contribute to creating a visual representation of an object presented to an observer. For instance, I examined the processes required for recognizing an apple (Chapter 5) and the visual representation of a horse (Chapter 6).

Through these examples, I aimed to demonstrate how the concepts and theories I analyzed and described can be applied in practice.

Importantly, my research methodology includes an element of theoretical synthesis. In the last part of the dissertation, drawing on the insights gained from conceptual analysis, critical review, case studies, and philosophical argumentation, I attempted to develop a modified and extended version of mechanistic explanation that can better accommodate the unique features of cognitive phenomena. In the theoretical modeling that I describe in the last part of the dissertation (6.4. *Towards “Cognitive Rich” Mechanistic Hierarchies in Explanatory Models*) I propose a new theoretical framework combining elements from different approaches (Marr's tri-level analysis, recursive decomposition, and Craver's mechanistic levels) to explain cognitive phenomena.

Finally, my entire thesis is permeated by a spirit of interdisciplinarity. The interdisciplinary integration attempted is reflected in the approach taken to address the research question, where philosophical reflection is combined with theoretical treatment of the scientific issues involved and the description of scientific explanatory approaches through selected case studies.

Structure of the Dissertation

The dissertation is divided into two parts: a more general first part that lays the theoretical foundations for biological sciences (*Part I: Foundations of Mechanistic Explanation: Levels, Decomposition, and Biological Complexity*), and a more specific second part that focuses on the phenomena investigated by cognitive science, which constitutes the research objective of the inquiry (*Part II: Exploring the Applicability of Mechanistic Explanations to Mental Phenomena*). I begin my investigation with two chapters that introduce the two fundamental concepts of the dissertation, which I consider essential for understanding the issue at hand: the concepts of levels and decomposition. In the next two chapters, I gradually focus on constitutive mechanistic explanations as described in the new-mechanistic framework, starting with the presentation of compositional explanations and culminating in the mechanistic model of explanation. The second part comprises two chapters. The first addresses the unique character of mental phenomena and focuses on the compatibility between the levels used within cognitive science and the mechanistic levels, ultimately arriving at a negative assessment. The second chapter concentrates on the discussion of integration between psychological and mechanistic explanatory models and the impasses this integration faces. It analyzes in depth the concept and utility of recursive decomposition in psychological models and concludes by proposing a solution for reconciling

mechanistic and psychological models into a unified, integrated form. Let us examine these in more detail in what follows.

In the first chapter (*Chapter One - Hierarchy and Levels in Biological Organization*) I initially present the concept of hierarchy and levels within the framework of biological organization. This is an essential first step toward a deep understanding of the philosophy behind the new-mechanistic explanatory framework. In the first part (*1.1. Laying the Groundwork: Introducing Hierarchy and Levels*), I present a common understanding of levels based on size, time scales, and intensity before addressing various notions of biological organization levels. I discuss different types of hierarchies and explain that the concepts of levels and hierarchy have a dual origin rooted in systems theory and mereology. I also outline the fundamental principles of hierarchy theory. The discussion proceeds in the second part (*1.2. Levels in Living Systems under a Biological Perspective*) with a detailed exploration of levels and hierarchy in the biological context, accompanied by the main critiques various accounts have received regarding their suitability for capturing the hierarchy of biological systems. I conclude this chapter by exploring levels that deal with complexity and ending with an introduction to the concept of mechanistic levels, setting the stage for further exploration in subsequent chapters.

In the second chapter (*Chapter Two - Decomposition as a Tool for Analyzing and Understanding Biological Systems*), I utilize ideas from systems theory, complexity science, and biology to present decomposition as a tool for analyzing and understanding complex biological systems. By exploring its role in understanding these systems, we can infer why decomposition is considered an integral part of the new-mechanistic approach to explaining their behavior. I begin by presenting the concepts of composition and decomposition, examining how these complementary processes relate to revealing the hierarchical organization of biological structures across multiple levels (*2.1. From Parts to Whole and Back Again: Exploring Composition and Decomposition*). This discussion sets the stage for a deeper investigation into the nature of complexity in biological systems (*2.2. Breaking Down the Complex: Decomposition in Biological Systems*). I then focus on two key aspects of biological organization (*2.3. Structural Principles and Organizational Dynamics in Complex Biological Systems*): Aggregativity vs Emergence and Interactivity vs Modular Organization. In the last part of the chapter, dedicated to the concept of near-decomposability (*2.4. Idealization and Near-Decomposability: Bridging Simplicity and Complexity*), I explore how this idea allows for idealized analysis of complex systems while still capturing both modular and emergent whole system properties.

In this chapter, it becomes clear that decompositional analysis is essential for biological systems and can work even when a system is not fully decomposable, which is usually the case in these systems, as they are characteristic examples of balancing between order and chaos. It is argued that analyzing a system through functional or componential decomposition can work, provided that the system is near-decomposable. It is proposed that many brain structures, as well as mental phenomena rooted in them, can be considered examples where near-decomposability applies. This observation leaves a window of applicability for explanations based on decomposition, even in systems that are not fully decomposable or modular in a strict sense. That is, while they have functionally specialized subsystems, they exhibit weak interactions between subsystems and use the same components for functions serving other modules. Consequently, in the second chapter, beyond highlighting the role of decomposition in analyzing biological systems and their behaviors, I support the tolerance of this tool in analyzing systems that are not totally modular and completely decomposable.

In the third chapter (*Chapter Three - Exploring Biological Mechanisms via Compositional Explanations*) which begins the discussion of compositional explanations, I initially present the dialogue that has led to supporting these explanations as more suitable for examining and analyzing biological systems and explaining their behaviors. I then introduce the concepts of decomposition and multilevel analysis as foundational elements for understanding biological mechanisms (3.1. *The Evolution of Explanatory Models in Biological Sciences*). The role of decomposition is further demonstrated through Gillett's (2016) work on scientific composition (3.2. *Gillett's Framework on Scientific Composition*), which offers a detailed account of how compositional explanations can be constructed and applied in biological contexts. I then, apply Gillett's framework in a detailed case study of an elbow movement (3.3. *Elbow Movement as a Model for Multilevel Biological Analysis*) to demonstrate how a complex biological system can be analyzed and explained across multiple levels of organization.

This chapter presents an argument in favor of compositional explanations. Through the example provided, we gain a deeper understanding of how this type of explanatory practice is applied in practice, with particular emphasis on how it unfolds across a hierarchy composed of multiple levels, where descriptions and explanations become increasingly detailed. Since the case study draws information from contemporary scientific textbooks, it simultaneously reflects the kind of explanations provided in the specific scientific field (physiology of movement, in this case). This

approach is generally encountered in many areas of biological sciences, as we will see from other cases examined briefly or extensively in the text.

In the fourth chapter (*Chapter Four - Decomposing Phenomena into Mechanisms in the New-Mechanistic Approach*), building upon the foundations laid in the previous chapters, I explore the new-mechanistic approach as a species of compositional explanations utilized to explain biological phenomena. I begin by tracing its development and application, particularly in the context of neuroscience, I present Craver's account of constitutive mechanistic explanation and address the crucial distinction between etiological and constitutive explanations (*4.1. Tracing the Roots of New-Mechanism*). In the second part of the chapter, I investigate any ambiguities in the definition and detection of a mechanism in the new-mechanistic sense and explore its boundaries and challenges through a discussion concerning borderline cases and the difficulties posed by highly complex biological systems (*4.2. What is and what is not a mechanism?*). In the final part, I present a detailed case study on spatial memory. It involves mice navigating through a Morris water maze and provides a practical demonstration of constitutive mechanistic decomposition into a hierarchy of levels of mechanisms (*4.3. Mechanistic Decomposition: A Case Study in Spatial Memory*). This example illustrates how the new-mechanistic approach integrates the concepts of levels, decomposition, and compositional explanation discussed in previous chapters.

In the exploration of the applicability of mechanistic explanations to mental phenomena which is the subject of the second part, I revisit the concept of levels, this time as it is understood in multiple ways within cognitive science. In the fifth chapter (*Chapter Five - The Multiplicity of Levels and Cognitive Science Explanations*) I examine the extent to which the concept of mechanistic levels is compatible with the various notions of levels employed in the explanations provided by the fields that constitute the multidisciplinary character of cognitive science. I begin by addressing a crucial question: What makes mental phenomena special? As I present the specific features of mental phenomena instantiated in biological systems, I emphasize what differentiates them from mere biological phenomena and what the implications are for the models employed to explain them (*5.1. The Unique Nature of Mental Phenomena in Explanatory Frameworks*). Here, I refer to intentionality and information processing as defining features of mental phenomena. In the second part of the chapter, I examine the role of multiple perspectives in cognitive science, providing an overview of diverse notions of levels used in mental phenomena research (*5.2. Navigating the Levels: The Role of Multiple Perspectives in Cognitive Science*) and present visual perception as a model for multilevel explanations in cognitive science. Next, aiming to inquire whether it is possible to combine

cognitive science explanations with mechanistic ones (*5.3. Reconciling Cognitive Science Explanations with Mechanistic Levels*), I revisit Craver's account, examining his ideas on mechanistic levels and the potential to apply them to mental phenomena. The chapter concludes with an important challenge in applying mechanistic models to cognitive explanations which regards the difficulties faced in incorporating informational and representational content into levels of mechanisms.

The final chapter (*Chapter Six - Cognitive Mechanisms: So Close, Yet So Far*) concludes the dissertation with three significant explorations. First, it provides a critical review of arguments and perspectives regarding the extent to which an integration of psychological and mechanistic explanations can be achieved in the mental realm. Second, it extensively presents the concept and application of recursive decomposition through the Information Processing Approach to Cognition. Third, it proposes a solution combining psychological and mechanistic explanatory models into a unified framework that bypasses the problems highlighted in the dissertation, specifically, the incompatibility of mechanistic levels with the levels used in cognitive explanations and the relationship between the hierarchies of physical and mental entities that emerge after the physical and informational decomposition of mental phenomena instantiated in biological systems.

I begin discussing the viability of applying mechanistic explanation to cognitive phenomena (*6.1. The Debatable Viability of Mechanistic Explanation in the Cognitive Realm*) by examining various perspectives on integrating psychological and neurobiological explanations, including arguments against the autonomy of psychological explanations and proposals for complementary or unified neurocognitive mechanistic models, the latter expanding on Marr's trilevel account as a potential framework for cognitive-mechanistic explanations. In the next section I address the challenge of incorporating mental properties into mechanistic hierarchies (*6.2. Is There Any Place for Mental Properties in Mechanistic Hierarchies?*). In the third part of the chapter, I introduce the concept of recursive decomposition (*6.3. Recursive Decomposition as a Key to Multilevel Cognitive Explanations*) as applied to information processing systems and present a case study in visual perception to demonstrate how this approach can be used to analyze cognitive phenomena. In the final part, I set out my proposal for a framework that could combine psychological and mechanistic explanations in a nearly integrated form (*6.4. Towards "Cognitive Rich" Mechanistic Hierarchies in Explanatory Models*). In this hypothesis I combine elements from recursive decomposition in information processing systems, Marr's trilevel account, and Craver's constitutive mechanistic explanation to create a more comprehensive model for explaining cognitive phenomena.

In a nutshell, the idea is the following: Instead of seeking a one-to-one correspondence between informational operations and physical processes, the solution I propose, maps operations to physical mechanisms. This approach maintains two separate but interrelated hierarchies: levels of mechanisms and levels of analysis. The framework begins with an undecomposed mental event, described by a mapping theory (a complex function that matches the incoming to the outgoing information) and a mental operation. To explain its physical (implementational) aspect, mechanistic decomposition is used to identify the spatiotemporally organized entities and activities constituting its physical mechanism. For the informational aspect, informational decomposition reveals sub-operations and mini-mapping theories. Further exploration can proceed, either by explaining the implementational aspect through additional mechanistic decomposition, or by analyzing the mental state through further informational decomposition. This solution allows for the methodological and theoretical interaction between psychological and neurobiological approaches while preserving the multiple realizability principle and avoiding the constraints of strict one-to-one mappings between informational and physical elements. The framework aligns with Marr's trilevel account and Craver's ideas about levels of mechanisms and realization, offering a comprehensive approach to understanding both the informational processing and physical implementation of cognitive phenomena.

It is important to note though, that this proposal is made with the reservation of the decomposability of the information processing systems to be explained. As has been discussed multiple times at various points in the dissertation, the "freedom" that near-decomposability provides in the use of decomposition in many complex systems allows for the formulation of such solutions even in controversial cases, as long as a balance between idealization and reality is maintained. Nevertheless, a thorough examination of this issue is certainly necessary, as well as inevitably extensive, and thus could not be included in the present inquiry to the extent that I estimate it would be necessary for providing a more definitive answer regarding the applicability of mechanistic explanation to mental phenomena.

Scope and Limitations

My dissertation focuses primarily on the philosophical aspects of applying mechanistic explanations to cognitive phenomena, with an emphasis on the conceptual and theoretical challenges this presents. While it draws upon examples and case studies from cognitive science and neuroscience, it does not aim to provide a thorough empirical analysis of specific cognitive

mechanisms. Instead, it uses these examples to illustrate and explore broader philosophical issues concerning levels and decomposition within specific explanatory models.

Also, my research is primarily concerned with human cognition and the explanatory frameworks used to understand it. While some of the insights may be applicable to artificial cognitive systems or animal cognition, these areas are not explicitly addressed within the scope of this inquiry. The rapidly evolving nature of cognitive science and neuroscience implies that new empirical findings or theoretical developments may emerge that could impact the arguments presented. It is very likely that these will also arise from the rapidly developing field of artificial intelligence, which is not in the scope of the present investigation but can be a source for updating our knowledge about mentality. I estimate that the framework I provided is flexible enough to accommodate such developments, although necessarily bounded by the current state of knowledge in these fields.

A significant limitation of this dissertation is the potential decomposability of the systems that are the target of my investigation. Although this issue has been touched upon in many instances, it should be emphasized that if the possibility of analyzing a system and a phenomenon through decomposition is questioned from the outset, then there is no point in discussing and demonstrating all the problems I have highlighted. These problems emerged with great caution regarding decomposability and under the assumption that these systems are near-decomposable. However, the concept of near-decomposability is nearly as complex as the systems to which it applies, and it is not a simple matter to fully understand and explore it in depth. It involves an investigation that most likely requires contributions not only from philosophy but also from complexity science, and needs to be studied both theoretically and empirically through various models. Regarding this crucial issue, the present dissertation maintains a pending attitude.

Lastly, it's important to acknowledge that the philosophical arguments presented in this dissertation, while thoroughly developed, are subject to ongoing debate and alternative interpretations. The conclusions drawn should be understood as contributions to an ongoing dialogue rather than definitive resolutions to the complex issues surrounding mechanistic explanations in cognitive science.

Conclusion and Acknowledgments

In this dissertation, I advanced several interconnected arguments that collectively contribute to a nuanced understanding of mechanistic explanations in cognitive science. My central thesis is that while the new-mechanistic framework offers valuable insights for explaining cognitive phenomena,

it requires significant modifications to fully capture the unique characteristics of mental processes. In essence, mechanistic explanation, as it has been developed at present, faces two significant challenges in its application to cognitive phenomena. First, mechanistic levels cannot be easily reconciled with the various concepts of levels encountered in cognitive science explanations, which reflect the informational and representational aspects of mental phenomena that are not captured by purely physical descriptions. Second, a significant challenge appears concerning how to relate and integrate the informational and physical hierarchies that emerge from the recursive decomposition of cognitive systems. Mechanistic explanations, properly extended and reconceptualized, can provide a unifying framework for integrating psychological and neurobiological accounts of cognition. Importantly, this integration does not entail reduction but rather a form of 'explanatory pluralism' that respects the autonomy of different levels and aspects of explanation while elucidating their interconnections.

I couldn't conclude this introduction without mentioning the particular challenges I encountered in its realization, as well as the people who stood by me, helped me, or inspired me with the depth of their thinking. First and foremost, I would like to thank my husband Michael, who, through our countless lengthy discussions on scientific and philosophical issues related to the nature of the mind, taught me how to think and instilled in me the passion to investigate in great depth the subjects I am dealing with. Moreover, he proved to be very efficient at our twins' babysitting, which made it possible for me to finish. With gratitude, I think of my professors in Cognitive Science who opened up intellectual horizons that I couldn't have imagined before. I am also grateful for the Department of Philosophy and History of Science, if nothing else, for its very existence and for being a place where anyone who wants to learn how to think can visit and stay.

In this department, I met people of very high intellect, and in every conversation with them, I discovered how much more I could think about. One such example is my supervisor, whose thinking has always set the bar high for what I wanted to achieve and thus gave me the impetus to try harder and investigate in greater depth. I would also like to make special mention of three professors in the department who provided significant support throughout my work on this dissertation. First, Eleni Manolakaki, who provided psychological support and helped me take the first step when, at a conference, she encouraged me by saying, 'You're on the right track, start writing it!' Then, Irini Skalioti, who inspired me more than anyone else when I was studying cognitive science and who always shows warmth and interest in what I do. Last but not least, Stavros Ioannidis, without whom I wouldn't have understood many of the things I deal with in the

field of Philosophy of Science. Through our discussions, our joint presence at conferences, the bibliography he constantly sent me, and his always calm and reassuring smile, he helped me understand in depth what I am investigating and complete my thesis.

Finally, I thank my two little twins, who systematically hindered every attempt I made to study and write, but simultaneously served as a motivating force to complete my thesis so that I could finally relax in their little arms without anxiety.

As we turn, now to the first chapter, I invite you to approach the following discussions with both critical rigor and open-minded curiosity. For in exploring how we explain the mind, we are, in a very real sense, exploring the nature of explanation itself, and perhaps even the limits of human understanding.

PART I

FOUNDATIONS OF MECHANISTIC EXPLANATION: LEVELS, DECOMPOSITION, AND BIOLOGICAL COMPLEXITY

In the first part of the dissertation, I aim to provide the foundations that will support the specialized investigation in the second part, which focuses on explaining mental phenomena instantiated in biological systems. It consists of four chapters that gradually lead to the presentation of the new-mechanistic explanatory framework, as applied to multi-level (constitutive) explanations and as it emerges through the methodological tool of decomposition. In the first chapter, I introduce the concept of hierarchical levels of organization in biological systems, aiming to demonstrate the necessity for multi-level explanations in many areas of biological sciences. In the second chapter, I address biological complexity and the use of decomposition to understand it, as initially conceived by Simon and Wimsatt (among others) and later by several new-mechanists. After presenting these two central axes of the dissertation, I focus on compositional explanations in the third chapter, drawing from the scientific composition model proposed by Gillet (2016). The presentation of compositional explanations, accompanied by an extensive case study on the biological function of the human motor system, prepares the reader for the introduction to the final chapter of the first part. This chapter concentrates on constitutive mechanistic explanations, primarily referencing Craver's (2007) work, and concludes with a case study on the mechanisms of spatial navigation in mice.

CHAPTER ONE - HIERARCHY AND LEVELS IN BIOLOGICAL ORGANIZATION

1.0. OUTLINE

In this initial chapter of the dissertation, I explore the concepts of hierarchy and levels, which are essential for a deep understanding of the philosophy behind the new-mechanistic explanatory framework. In the first part (1.1. *Laying the Groundwork: Introducing Hierarchy and Levels*), I present a common understanding of levels based on size, time scales, and intensity before addressing various notions of biological organization levels. I discuss different types of hierarchies and explain that the concepts of levels and hierarchy have a dual origin rooted in systems theory and mereology. I also outline the fundamental principles of hierarchy theory. The discussion proceeds in the second part (1.2. *Levels in Living Systems under a Biological Perspective*) with a detailed exploration of levels and hierarchy in the biological context, accompanied by the main critiques various accounts have received regarding their suitability for capturing the hierarchy of biological systems. I start by examining levels in science and nature (1.2.1.), then address levels in neurobiology (1.2.2.), and introduce extended levels as well as some other special accounts centered on features like organism-environment interactivity and time scales (1.2.3.). I conclude this chapter by exploring levels that deal with complexity (1.2.4.), starting with Simon's ideas on hierarchy and complexity and ending with an introduction to the concept of mechanistic levels, setting the stage for further exploration in subsequent chapters.

1.1. LAYING THE GROUNDWORK: INTRODUCING HIERARCHY AND LEVELS

The world around us is organized into higher and lower levels. Some of these levels are ordered by size, as seen with microbes, animals, and planets. Others are defined by time scales, such as centuries, years, months, hours, or seconds. There are also intensity scales that classify phenomena as strong, mild, or weak. One can imagine several other such categorizations. Entities at each level often interrelate according to level-dependent constraints. For instance, large objects like planets are usually associated with long-lasting events, such as Earth's annual revolution around the sun. Additionally, things of different sizes often form hierarchies where lower levels are embedded

within higher ones. For example, land comprises continents, which consist of mountains and plains, which in turn contain forests, rivers, and lakes. These are made up of soil, trees, water, fish, and animals, and this analysis continues down to the microscopic level and beyond. This rough understanding of levels' hierarchy aligns with the image of nature familiar to most observers.

Perhaps the most widely used application of levels and hierarchy is in biological organization. A common classification lists more than ten levels, ranging from the smallest to the most extensive in size. Some classifications, based on the cell being the fundamental unit of life, begin at the *cellular level*. The standard biological organization scheme (Solomon, Berg, & Martin, 2011) includes even smaller entities, starting from the *acellular level* (atoms and molecules), progressing to the *pre-cellular* (biomolecular complexes), and the *subcellular* (organelles, functional groups of biomolecules, and biochemical reactions and interactions). Moving up from the cellular level, we encounter the *tissue level*, comprising aggregates of cells with similar functional or anatomical properties. Collections of tissues that function as a whole form *organs*, and multiple organs working together for a specific bodily function constitute an *organ system*. The next level is the *organism*, composed of various organ systems working in harmony to maintain homeostasis and ensure the animal's survival. Above the organism level are the *ecological* levels, including *populations* (groups of organisms of the same species) and *communities* (interacting populations of different species). Within the ecological domain, we also find *ecosystems* (groups of organisms in an environment), *biomes* (extensive geographical areas with similar climates and ecosystems), and finally, the *biosphere*, encompassing all living beings and their physical environment on Earth.

Beyond the illustrative paradigm discussed above, the concepts of hierarchy and levels are employed in various ways, for diverse purposes, and across a wide range of contexts. Indeed, these concepts are prevalent in both natural and artificial realms. We can categorize hierarchies into different types based on various criteria: in terms of their content and dimensions, they can be classified as spatial or non-spatial, structural or functional, living or non-living, as well as political, social, religious, economic, ecological, and physical hierarchies (Wu, 2013). This inquiry will not address the notion of hierarchy in science levels, which refers to classifications of scientific entities such as descriptions, theories, fields, and research programs. Additionally, I will not delve into the concept of realization levels at this stage. I will revisit this topic in Chapter 5 when discussing mental phenomena, where the consideration of realization levels will supplement my analysis.

The concept of levels and hierarchy used in this work has dual origins. First, it stems from systems theory, particularly the analysis of complex systems. In this context, it serves as a crucial element for understanding structures of organization, facilitating both explanation and prediction. Here, we are primarily concerned with levels of organization. Second, I will employ the mereological concept of hierarchy and levels, where part-whole relationships define the levels. In the following section, I will present several approaches to defining levels and hierarchy, with a focus on biological, neurobiological, and cognitive domains. Through these discussions, I aim to provide a thorough understanding of how the concept of levels' hierarchy is applied in biological systems. This groundwork will then allow us to explore the notion of system decomposition, which is closely tied to the concept of levels' hierarchy.

To begin this discussion, we must briefly touch on the fundamental principles of *hierarchy theory*, which was developed in the *general systems* research field and has its roots in Herbert Simon's work (Simon, 1962). Allen (2006) provides a concise summary of these principles: Hierarchy is defined as an arrangement of parts within a whole, characterized by ordered, asymmetric relations between different levels. Each level contains entities with properties that are characteristic of this level, and an entity may belong to multiple hierarchies depending on its properties. Ahl and Allen (1996) and Allen (2006) identify two main types of levels: The first regards levels of organization that fit into a hierarchy by virtue of a set of *definitions*; that is, they are defined by specific criteria rather than scale. For instance, in biology, a population is classified at a higher level than an individual organism based on its definition (a large group of organisms), not its size. Thus, a single large organism would be at a lower level than a population of microscopic parasites living on its skin. The second type regards *levels of observation* based on relative scaling factors, such as spatiotemporal criteria like size. The ordering of levels is determined by several criteria, which may be applied individually or in combination. Ahl & Allen (1996) propose five criteria for distinguishing a higher level from a lower one: These are: (1) the higher level provides context for the lower level, (2) the higher level constrains the lower level, (3) the higher level operates more slowly or at a lower frequency than the lower level, (4) the higher level consists of entities with greater integrity and stronger bonds, and (5), the higher level contains and is composed of lower levels.

The criterion of containment is crucial for our inquiry as it distinguishes between nested and non-nested hierarchies (Ahl & Allen, 1996). In a nested hierarchy, containment is always a defining feature of level order, alongside other criteria. For instance, animal organs are classified at a lower level than the organism, not only due to their size but also because the organism contains them.

However, when examining a specific individual, say George, it becomes irrelevant to mention that Peter's organs are at a lower level than George's organism, as George does not contain them. Similarly, when studying an organism's cardiovascular system, it's not pertinent to state that alveoli are at a lower level than the aorta merely because of their smaller size. Alveoli belong to the respiratory system, not the cardiovascular system. They would only be considered at a lower level if size were the sole criterion, which would indicate a non-nested hierarchy. Ahl & Allen (1996) argue that non-nested hierarchies are fundamentally different from nested ones, having a more general nature. This study focuses on mechanistic explanations, where hierarchies always refer to nested levels with part-whole (or compositional) relationships (Craver, 2007; Bechtel, 2008).

Ahl & Allen (1996) also highlight an intriguing aspect of deep hierarchical structures: as we descend levels, lower ones exhibit fewer degrees of freedom and simpler behavior. Moreover, hierarchical depth often indicates a sophisticated organization. We can illustrate this concept through the study of an organism's kinetic behavior using the notion of *degrees of freedom* (DOF) (Oatis, 2009). The complexity of an organism's movement depends on the range of motion allowed by its skeletal joints. We can describe movement relative to a coordinate system. Some joints permit movement in three spatial directions, thus having three DOFs, while others allow movement in only one direction, with one DOF. For example, the upper extremity's joints collectively have about 30 DOFs (an approximate figure depending on the level of detail considered). In contrast, the index finger alone has 4 DOFs, corresponding to its joint movements. Clearly, the index finger's motor behavior is less complex than that of the entire upper extremity, which exists at a higher level due to both its size and the inclusion of the finger. Therefore, breaking down upper extremity movement into lower-level components (joints) and functions (joint movements) serves to reduce complexity and facilitate the explanation of the behavior.

The mereological concepts of hierarchy and levels in material compositional hierarchies share many similarities with the levels of organization notion in complex systems discussed earlier. In essence, when dealing with a nested hierarchy, a systemic definition inherently incorporates mereological aspects. Mereology focuses on parthood relations, specifically examining how parts relate to the whole and to each other within that whole. The term *parthood* broadly indicates that something is a component or portion of an entity. However, this general definition requires further elaboration to precisely determine the nature of the parthood relation (Varzi, 2019). For example, a part may belong to a set of either homogeneous or heterogeneous components that are either spatially connected or separated. Additionally, an element might be considered part of a whole due

to its functional role, its material composition, or both. These nuances help differentiate the specific character of parthood relations. Clear examples of material constitutive parthood include the relationship between clay and the statue it forms or the “A” key on a computer keyboard. However, applying this concept becomes more complex in cases like a goalie being part of a football team. This example differs significantly, and it's debatable whether a team truly constitutes a mereological whole (Varzi, 2019). Considering these complexities, in his taxonomy of levels, Craver (2007) argues that mereology is a relatively narrow concept that often overlooks the relations held among parts. He categorizes it as a specific case of simple parthood arising from compositional relations. Craver emphasizes that in neurobiological systems, a purely mereological approach to levels often fails to capture the complex relations between components that form a whole, risking the oversimplification of complex systems into mere aggregates. Consequently, Craver opts for the more comprehensive term *compositional levels* instead of mereological levels.

1.2. LEVELS IN LIVING SYSTEMS UNDER A BIOLOGICAL PERSPECTIVE

This section reviews some of the most well-known level accounts to provide a general understanding of the meaning and application of levels in biological systems. This overview will serve as a foundation for comprehending how we decompose these systems for explanatory purposes. Many of these accounts have been critically examined by mechanistic philosophers regarding their suitability for explaining biological phenomena. I will focus on critiques concerning the appropriateness of these accounts in analyzing and explaining biological, neurobiological, and cognitive systems and their associated phenomena.

I begin this review with a brief discussion of the *layer-cake model* (Eronen & Brooks, 2018), which is based on Oppenheim and Putnam's (1958) ideas about the unity of sciences. I then proceed to examine the neurobiologically focused account of Churchland and Sejnowsky (1992). As the review progresses, I present more sophisticated accounts that attempt to capture various aspects of the complexity inherent in biological and cognitive phenomena. These include Gibson's notion of levels, which is part of his ecological psychology approach (Gibson, 1986), and Newell's (1994) account, which considers the role of time in defining levels in biologically instantiated cognitive systems. The review concludes with Simon's ideas about hierarchy and complexity (Simon, 1962), which have inspired, among others, Wimsatt's account of biological systems levels (Wimsatt, 1976; 1994). Simon's and Wimsatt's accounts are closely linked to their ideas about decomposition, and

therefore, they will be further explored in the next chapter, which deals with decomposition in complex biological systems. The constructive critique that Wimsatt's levels received, along with modifications proposed by proponents of the new-mechanistic movement, led to the idea of *levels of mechanisms*. This idea will be briefly introduced in this section and discussed in more detail in the fourth chapter, which focuses on the New-Mechanism movement's approach to explaining biological phenomena. It will also be addressed in the fifth chapter, which examines levels in cognitive science research.

1.2.1. Levels in science and nature

Levels have been widely employed not only for analyzing and explaining complex systems but also in other scientific contexts. One of the most notable applications of this notion relates to the pursuit of scientific reduction, which aims to achieve a unified science through the reductive unification of concepts and laws to those of the most fundamental entities (Cat, 2017). The underlying idea is to identify bottom-level entities and theories that are presumed to be explanatorily sufficient, potentially rendering top-level entities or scientific theories ultimately redundant. Within this context, the concept of levels in sciences and nature is integral to a reductive approach. Oppenheim and Putnam (1958), pioneers of the scientific reduction movement, proposed a mereological system of levels that would provide a basis for ordering entities to facilitate the micro-reduction of higher-level entities to lower ones. In this system, they suggested a set of principles, which are as follows:

- 1) There must be several levels, 2) The number of levels must be finite, 3) There must be a unique lowest level (...), 4) Anything of any level except the lowest must possess a decomposition into things belonging to the next lower level. (...), 5) Nothing on any level should have a part on any higher level, and 6) The levels must be selected in a way which is "natural" and justifiable from the standpoint of present-day empirical sciences. (...)
- (Oppenheim & Putnam, 1958, p. 9).

Consequently, they propose a list of six levels that satisfy the above conditions, arranged from highest to lowest: social groups, multicellular living things, cells, molecules, atoms, and elementary particles. These levels correspond not only to natural entities but also to the sciences that study them, thus serving a role in theoretical reduction. However, this account, which defines levels in a rather rigid manner by segmenting nature into distinct layers, appears to oversimplify or hastily resolve the complexity of many phenomena, particularly in biology. This oversimplification is

unsurprising, given that the account was developed with the intention of facilitating scientific reduction.

This approach has faced widespread criticism. The main points of contention focus on the oversimplification inherent in suggesting that the world can be neatly divided into six levels. Critics argue that this system has significant gaps and fails to include entities that could exist at different levels, potentially leading to inconsistencies. Craver (2007) particularly emphasized that this account is inadequate for studying and explaining neurobiological phenomena, especially in light of the methodologies employed in contemporary neuroscience.

1.2.2. Levels in neurobiology

Churchland and Sejnowski in their seminal work on computational neuroscience (1992), propose an approach that considers entities crucial to neurobiological explanations. They examine the concept of levels as commonly used in neurobiological and psychological research, aiming to understand its application in various contexts. To achieve this, they distinguish three different notions of levels. (1) *Levels of analysis*: These are conceptual levels addressing how the brain performs specific functions. Churchland and Sejnowski cite Marr's trilevel approach (Marr, 1982) as an example, which comprises computational, algorithmic, and physical implementational levels. While the relation between these levels is not compositional (Craver, 2007; Bechtel, 2008; Zednik, 2017), they are generally considered realization levels (Craver, 2007). (2) *Levels of processing*: This non-mereological notion is based on a process's position in a processing sequence. The closer a process is to sensory input (the periphery), the lower its level. For instance, retinal visual signal processing is considered lower-level than processing in the lateral geniculate nucleus (LGN), while processing in the primary visual cortex (V1) is higher still. This taxonomy follows a temporal sequence, often with cause-and-effect relations between levels. (3) *Levels of organization*: This mereological notion is primarily based on size scale and includes both anatomical and physiological levels. Anatomical levels refer to “organized structure at different scales: molecules, synapses, neurons, networks, layers, maps, and systems” (Churchland & Sejnowski, 1992, p. 19). Physiological levels address the function of these organized elements, answering the question, “What does that organization of elements do?” (Churchland & Sejnowski, 1992, p. 19).

Churchland and Sejnowski focus primarily on anatomical levels, which they explore in greater detail in their work. They are the following seven, arranged from highest to lowest: At the top is the *level of brain systems*, encompassing large brain areas that correspond to functions such as long-

term memory or specific sensory modalities, as well as less clearly defined functions. For instance, a system comprising the hippocampus, thalamus, frontal cortex, and basal forebrain nuclei is strongly associated with long-term memory processes. The second level consists of *topographic maps*, which are well-defined organizational structures found in sensory and motor systems. These maps comprise groups of neurons arranged to reflect the structure of external stimuli or kinetic activation patterns within the brain. They often form processing hierarchies; for instance, the visual system represents the external world at various stages of the visual pathway through topographic maps. The third level includes *laminae (layers) and columns*. Third, the level of laminae or *layers and columns*. A lamina is a neuronal layer characterized by a highly regular pattern of efferent and afferent connections. Columns, on the other hand, are vertically organized cells with numerous local connections and similar response properties, such as the ocular dominance columns found in the primary visual cortex (V1). The fourth level, *local networks*, refers to populations of thousands of neurons and approximately a billion synapses with a particularly complex organization. These networks are likely responsible for functions such as object representation or the computation of color constancy. The fifth level comprises *neurons*, the functional and anatomical units of the nervous system. Neurons exist in various types and possess a range of specialized properties. The sixth level consists of chemical *synapses*, the structural units where neurons communicate with each other. These synapses occupy areas on the order of a few square micrometers. Finally, at the lowest level are *molecules*, such as the membrane proteins that form ion channels or the molecules that make up the second messenger system.

As evidenced by the examples in the previous paragraph, Churchland and Sejnowski attribute specific functions to each level. They propose that the function performed by a specific anatomical structure and the structure itself (the neurobiological substrate) represent two different perspectives of the same organizational level. Consequently, the seven anatomical levels described roughly correspond to related physiological levels. It's worth noting that they refrain from suggesting an exact match between anatomical and physiological processing levels. The proposed physiological levels include ion movement, channel configurations, excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs), action potentials, and evoked response potentials. For higher anatomical levels, such as systems or networks, they leave the corresponding physiological levels open-ended.

Churchland and Sejnowski's classification scheme undoubtedly captures the specific features of complex biological systems more accurately than Oppenheim and Putnam's (1958) account, as it is

grounded in the description and study of neurobiological functions. However, its application in formulating explanatory models has faced criticism, particularly from proponents of the new-mechanistic movement in biological explanations (Craver, 2007; Bechtel, 2008). The main critique is that this ordering of levels is based primarily on size, neglecting the compositional relations between levels. These compositional relations are considered essential for explanations appropriate to the biological domain. For instance, according to this taxonomy, a liver cell would be classified at the same level as a neuron despite being parts of different structures and phenomena. Their classification at the same level based solely on size is not explanatorily relevant. New-mechanists argue that levels used in explaining phenomena should be local rather than global (Craver, 2007; Bechtel, 2008). Furthermore, Churchland and Sejnowski's account is limited in its applicability, as it is exclusively suited for analyzing brain-bounded systems, being restricted to the scale of structures within the brain. This limitation renders it incompatible with explanations that incorporate factors external to the organism, such as ecological or social environments.

1.2.3. Extended levels

The latter issue is not a concern for the next two approaches presented below, which propose levels that extend beyond brain boundaries to include the body and even the environment. The first approach relates to James J. Gibson's *ecological psychology* (Gibson, 1986). Gibson proposed an original and rather controversial account of levels in nature, which he used as a foundation to develop ideas in ecological psychology concerning how living organisms interact with their environmental niches. Gibson's conceptual framework differs significantly from that used in traditional cognitive science and biology, viewing cognition as a genuinely externalistic, situated, and enactive phenomenon. With this perspective, he suggested a hierarchy of levels in nature that takes into account both the size scale and the interactivity profile of entities at each level. Central to this hierarchy is the role of an entity with psychological attributes -the observer- who both performs the classification of levels and acts within them. The animal-environment interplay is the key determining factor in Gibson's account of levels in nature. These levels are descriptions of the world, defined according to the distinction between animate and inanimate beings. The focus is on the habits of animate beings and the complementarity of the animal-environment pair. While world descriptions in physics deal with extreme values of size, such as fundamental particles at the microscopic scale and galaxies at the macroscopic scale, neither of these constitutes an environment for living, animate beings. Instead, life descriptions reside at an intermediate level,

with sizes ranging from millimeters to kilometers. Within this range, Gibson defines biological hierarchy levels that concern beings living in the described environment.

Gibson describes several environmental levels: The *level of the order of kilometers*, encompassing mountains and hills; the *level of the order of meters*, including rocks, canyons, caves, and trees; and the *level of the millimeters' order*, comprising pebbles, crystals, soil granules, leaves, and grass. He also proposes a microscopic level, which includes plant cells. These levels form nested hierarchies; for instance, a canyon is nestled within a mountain, and a tree is nestled within a canyon. Each environmental level is associated with organisms that inhabit it based on the specific affordances it offers. An affordance is an action possibility provided by the environment to an organism, dependent on the organism's physical attributes and capabilities. In Gibson's view, levels are co-defined by the environment and the living beings that occupy and interact within an ecological niche. This approach, which emphasizes the interaction between living entities and their environmental niches, is a fundamental aspect of Gibson's ecological psychology. It formed the basis for his ecological theory of visual perception, which stresses the complementarity between observer and observed, challenging many theoretical and methodological principles of traditional cognitive neuroscience. This perspective is included in this review because it underscores the importance of interactions, a theme that will also be evident in Wimsatt's approach.

Allen Newell's (1994) approach is particularly noteworthy as it considers not only size-defined levels but also temporal levels of organization. Similar to Gibson's account, interactions play a crucial role in Newell's framework. He defines a system level as "a collection of components that are linked together in some arrangement and that interact, thus producing behavior at that system level" (Newell, 1994, p. 117). Newell's account primarily focuses on biological cognitive systems. Drawing inspiration from Simon's earlier ideas about hierarchy and levels, Newell grounds his work in systems theory and computer science. He conceptualizes the human mind as an intelligent control system composed of multiple levels of subsystems. Newell's proposed account of levels is influenced by Simon's analysis of natural hierarchies and his ideas about the necessity of levels for the development of stable forms in nature. Importantly, Newell views these levels as abstractions, describing them as "alternative ways of describing the same system, each level ignoring some of what is specified at the level beneath it" (p. 118).

Newell's categorization of levels is fundamentally defined by two parameters: size and duration. As he states, "As one moves up the hierarchy of system levels, size increases and speed decreases."

(Newell, 1994, p. 119). Size plays a crucial role in determining level order. Consider a lower level with K components, each of size S , composing an entity at the upper level. The size of this upper-level entity will be at least KS or potentially larger when accounting for the spatial distribution of its components. This principle applies recursively to each subsequent level. Newell demonstrates that each higher hierarchical level necessarily consists of entities geometrically larger than those at lower levels, thus establishing size as a determinant of a level's hierarchical position.

A similar expansion in duration applies to processes across hierarchical levels. Higher-level processes, constituted by lower-level processes, invariably require more time for completion. If a lower-level subsystem needs a specific time to generate a response to an input, then a higher-level system composed of multiple such subsystems will require more time to produce its outputs. The precise calculation of this duration is complex and influenced by the arrangement and interconnections of the subsystems. Importantly, it is logically impossible for two different system levels to operate on the same time scale, as this would imply a lack of interaction between the subsystems, negating the concept of a higher-level system. In summary, ascending the hierarchical scale correlates with increases in both size and duration. Newell argues that in hierarchies of units, more abstract units are necessarily larger and slower than their more concrete counterparts.

Newell's (1994) framework of temporally defined levels of organization offers a comprehensive view of processes related to human action. This hierarchical model is structured such that as one ascends the levels, both the duration of processes and the size of participating entities increase. Durations are measured in seconds on a logarithmic scale, with each level lasting ten times longer than the one below it. Newell organizes these levels into several *bands*, each corresponding to different phenomenal worlds and described by distinct theories. The hierarchy begins with the *biological band*, which comprises three levels. At its core is the neuronal level, where individual neurons operate at durations of approximately 1 millisecond, with cell bodies measuring about 0.01 millimeters. Above this lies the neural circuits level, functioning at around 10 milliseconds, while below is the cell organelles level, operating at 100 microseconds. Moving upward, we encounter the *cognitive band*, which encompasses three levels of increasing complexity. The lowest level concerns any deliberate act, the middle corresponds to a cognitive operation, and the highest represents a unit task. Above the cognitive band lies the *rational band*, with durations ranging from minutes to hours, followed by the *social band*. Newell also proposes two even higher-level bands with significantly broader durations: the *historical* and the *evolutionary*. However, he acknowledges uncertainty about whether these truly constitute separate bands within his framework. Despite the

expansive nature of this model, Newell's primary focus remains on the lower levels of organization, particularly those within the biological and cognitive bands.

Ballard et al. (1997) refined Newell's time scales to align with experimental observations, proposing a revised hierarchy of temporal bands that organize human computation. At the top is the *cognitive level*, corresponding to actions like dialing a phone number, which typically lasts 2-3 seconds. Below this is the *embodiment level*, encompassing actions such as eye movements, with durations of about 0.3 seconds. The *attentive level* follows, involving processes like noticing a stimulus, which occurs within 50 milliseconds. Ballard et al. further delineate two neural levels: a higher one corresponding to neural circuits (e.g., lateral inhibition) operating at 10 milliseconds and a lower level representing basic signals like individual neuron spikes, occurring within 1 millisecond. This temporal classification provides an alternative framework for organizing levels, focusing on process durations rather than the physical entities involved. By emphasizing computational processes over the structures in which they are implemented, this approach allows for the identification of explanatory levels that may extend beyond the body. Ballard et al. proposed this scheme specifically to elucidate mental phenomena involving distributed processing between cognitive agents and their environments. Their model offers a more nuanced understanding of the temporal dynamics underlying human cognition and behavior, bridging the gap between neural activity and higher-level cognitive processes while acknowledging the role of the external environment in shaping mental phenomena.

1.2.4. Levels that deal with complexity

Newell's conceptualization of levels is particularly relevant to our inquiry into cognitive systems, which are active systems with entities, interactions, and activities spanning space and time. By using size and duration to distinguish levels, Newell provides a framework that addresses both the physical entities in space and the processes in which they are involved. Space and time are crucial components for studying the organization of entities within an active complex system. Simon (1998) further emphasizes the importance of time scales in explanations, distinguishing between *descriptive* and *explanatory* scientific theories. Both types of theories connect dependent and independent variables, but explanations aim to identify the cause of a phenomenon, often dealing with theoretical terms and variables that are not directly observable due to their smaller size and temporal scale. According to Simon, explanations generally involve finer temporal resolutions. This aligns with the common practice of decomposing phenomena into parts and processes at lower

scales and levels to explain them. Consequently, the deeper the explanation, the smaller the duration of the processes involved. This parallels the role that size plays in explanatory depth. Simon illustratively compares explanations to a child's recursive sequence of "why" questions, potentially leading to an infinite regress of explanations at ever-smaller size scales.

It is important to consider the role of Simon's ideas in shaping innovative trends in general systems theory and the study of complex systems. Prior to his inspirational paper "The Architecture of Complexity" (1962), the concept of hierarchy and levels was primarily associated with reductionist approaches, and complexity science was heavily influenced by chaos theory. Moreover, there was a strong connection between general systems theory and formal hierarchies. Formal hierarchies refer to organizational structures with a pyramid-like arrangement, where a leader is at the top and subordinate subsystems occupy lower levels. Simon's propositions, however, do not pertain to this type of formal hierarchy; rather, his focus is on compositional hierarchies. He defines a *hierarchic system* as "a system that is composed of interrelated subsystems, each of the latter being, in turn, hierarchic in structure until we reach some lowest level of elementary subsystem" (Simon, 1962, p. 468). Simon does not address entities that can be divided into an indefinitely large number of subsystems, such as the numerous carbon atoms in a diamond's crystal structure. These are considered *flat hierarchies* and are less intriguing in terms of complexity compared to hierarchies with a moderate number of subsystems at each level.

His significant contribution was the assertion that "complexity frequently takes the form of hierarchy" (Simon, 1962, p. 468). This proposition has two main points: complex systems have structure, and one of the primary structures they exhibit is hierarchical. He further argues, without teleological implications, that there is an evolutionary advantage in developing complex hierarchies because it is more economical in terms of time and stability for a system to develop intermediate stable forms during its evolution. Examples of such stable forms in biological systems include single-celled organisms and macromolecules. Simon also notes that in complex natural systems, such as the mind, hierarchical structure reduces operational complexity to that of a much simpler system (Simon, 1995). This is because each subsystem has a specialized function, and they work together to complete an entire system process. Regarding the application of hierarchical levels in cognition, Simon (1998) views the mind as a physical symbol system; here, the computer metaphor has played a crucial role in shaping traditional cognitivist theories in cognitive science. Similar to compositional hierarchies comprising spatially extended entities like macromolecules, membranes, and neurons, an information-processing hierarchy can be conceptualized. Indeed, the process of

breaking down a problem into smaller components can reduce complex computations to a sequence of simple steps, which is fundamental to the inspiration behind early computing algorithms. This concept also underpins the *Physical Symbol System Hypothesis* as a representation of thinking.

Simon's ideas significantly influenced the development of accounts applying complexity theory principles to biological systems. Wimsatt, a prominent philosopher, expanded on Simon's inspirations to conduct an in-depth study of hierarchy concepts in biology (Wimsatt, 1976; 1994). He developed an extensive and somewhat complex account of nature's levels, primarily encompassing real-world entities: "I will assume that being at a given level is a property primarily of things in the world: phenomena, objects, properties, processes, causes and effects, etc., and derivatively of linguistic things relating to them: descriptions, law-statements, theories, predicates, etc." (Wimsatt, 1976, p. 215). Wimsatt emphasizes that his use of "levels" only derivatively refers to linguistic or theoretical constructions, as his account is primarily ontological. In the case of compositional levels, they refer to real and robust entities. According to his robustness criterion, "things are robust if they are accessible (detectable, measurable, derivable, definable, producible, or the like) in a variety of independent ways" (Wimsatt, 1994, p. 2). For Wimsatt, the robustness of entities and parts at each level is a prerequisite for compositionality in a hierarchical system.

Ontologically, levels of organization in nature represent networks of causal interactions organized into patterns under certain conditions, such as size or spatial containment. Wimsatt considers causal relations the primary working matter of the world and, on that basis, offers two other organizational schemes beyond levels: *perspectives* and *causal thickets*. Perspectives are "quasi-subjective (or at least observer, technique or technology-relative) cuts on the phenomena characteristic of a system, which needn't be bound to given levels" (Wimsatt, 1994, p. 17). They refer to much higher or very low levels with more degrees of freedom and many overlapping and interrelated properties with neighboring levels. Perspectives and causal thickets are methodological tools suggested by Wimsatt for studying systems characterized by high complexity, context dependency, low modularity, and low regularity. These systems cannot be studied based on well-defined levels of organization since they have a more diffuse character, are less robust, and their interactions can *bridge* several neighboring levels of different size scales. By "quasi-subjective," Wimsatt refers to a character that resembles a *point of view* but is not exactly the same.

Causal thickets come into play when organizational complexity is too great to be handled by a single perspective, and there are disorder and boundary ambiguities such that level claims cannot be made without serious dispute. However, levels are much more robust than perspectives and causal thickets. As he states, "Whatever levels are, they are at least major vertebrae among the joints of Nature" (1976, p. 237). This ontologically profound statement is based on the metaphor that "the aim of science is to cut Nature at its joints"¹ (p. 237). Wimsatt acknowledges Simon's influence on his ideas, stating that his almost Kantian view of organizational levels would be shared by Simon: "I urge a view that Simon would share: that levels of organization are a deep, non-arbitrary, and extremely important feature of the ontological architecture of our natural world, and almost certainly of any world which could produce, and be inhabited or understood by, intelligent beings" (Wimsatt, 1994, p. 6).

So, Wimsatt's levels are very much levels of nature; nevertheless, the principles by which he specifies them are more statistical than universal, deriving from the character of phenomena in complex systems. In his account, he offers mainly a prototype of levels instead of a definition. However, he describes levels as "local maxima of predictability and regularity" (Wimsatt, 1976, p. 238) and illustrates this with a graph where the horizontal axis represents the size of entities, and the vertical axis represents the regularity and predictability of their interactions. Although interactions can be found between entities of different sizes, there is a much higher density of interactions within the same size scale. This results in a relatively tight clustering of entities at a given size level. This definition is not particularly rigorous, as his approach is not intended to give a precise classification of levels but instead proposes organizational principles based on which the physical world is structured. In the case of levels of composition, these are principles applying to how matter is organized. At higher scales, far from our anthropocentric perspective, where properties of levels like size become less well-defined, the compositionality principle weakens.

Bechtel (2008) and Craver (2007), prominent proponents of the new-mechanistic movement in the philosophy of science, consider Wimsatt's account regarding levels in nature as one of the most extensive. However, regarding their utilization for explanatory purposes, they criticize his overestimation of the role that size plays in determining levels. In particular, Wimsatt's proposal

¹ This, in turn, is a paraphrase of Plato's (Phaedrus): "The second principle is that of division into species according to the natural formation, where the joint is, not breaking any part as a bad carver might" (Plato, Phaedrus: 265d - 266a).

that entities of the same size tend to interact more frequently and predictably has been challenged with several counterexamples. They propose a more local and process-based account that correlates working entities and activities (or operations) and classifies them into levels according to their role in a mechanism responsible for a given phenomenon. This account is much more local in scope and avoids problems such as a possibly unclear and ambiguous correspondence of several parts to a particular whole².

1.3. SUMMARY AND CONCLUSIONS

In this chapter, after an introductory presentation of the concept of levels and hierarchy through the lens of systems theory and mereology, I aimed to demonstrate the importance of levels to biological inquiry while highlighting ongoing debates about their proper characterization. My introduction to the concepts of hierarchy and levels focused mainly on their application to complex biological systems. Beginning with common understandings of levels based on size, time scales, and intensity, I then reviewed several influential accounts of levels, including the *layer-cake model*, Churchland, and Sejnowski's *neurobiological levels*, Gibson's *ecological levels*, and Newell's *temporal levels*. Through this exploration, it became evident that while these various approaches capture important aspects of biological organization, they fall short of addressing the full complexity of biological phenomena. This limitation underscores the challenge of developing a comprehensive framework for understanding levels in biology that can accommodate the complex and multi-faceted nature of living systems.

In the final part of the chapter, I examined Simon and Wimsatt's approaches, which more directly engage with complexity. I discussed Wimsatt's inspiration from Simon's ideas to develop an ontological account of levels understood as robust patterns of causal interactions. This perspective represents a significant step forward in dealing with the complexity of biological systems, offering a different and deeper understanding of how various levels of organization interact and influence each other. While acknowledging the comprehensiveness of Wimsatt's account, I also presented critiques from new-mechanistic philosophers regarding its overemphasis on size as a criterion for levels taxonomy. This criticism highlights an ongoing tension in the field between accounts that

² I shall present it in more detail in chapters 4 and 5, which deal with mechanistic explanations in neuroscience and cognitive science domains.

prioritize structural features (like size) and those that emphasize functional relationships and causal interactions.

The chapter concluded by introducing the mechanistic levels approach, which defines levels more locally based on compositional relations within specific mechanisms. It offers a more flexible and context-sensitive framework for understanding levels in biology. By focusing on the specific constitutive mechanisms underlying biological phenomena, it may be better equipped to capture the dynamic, interconnected nature of biological systems. This will be further demonstrated in the next chapters of the present dissertation.

In reviewing these various perspectives on levels and hierarchy in biology, several concluding remarks emerge. First, the concept of levels is fundamental to our understanding of biological systems, providing a crucial framework for organizing and interpreting complex phenomena. Second, there is no single, universally accepted account of levels in biology; rather, different approaches offer valuable insights into different aspects of biological organization. Third, as our understanding of biological complexity has grown, there has been a shift towards more flexible, context-dependent conceptions of levels that can better accommodate the diverse and dynamic nature of living systems. This introductory chapter sets the stage for extending the discussion into the concept of decomposition as well as exploring compositional and mechanistic explanations in subsequent chapters.

CHAPTER TWO - DECOMPOSITION AS A TOOL FOR ANALYZING AND UNDERSTANDING BIOLOGICAL SYSTEMS

2.0. OUTLINE

In this chapter, I utilize ideas from systems theory, complexity science, and biology to present decomposition as a tool for analyzing and understanding complex biological systems. By exploring its role in understanding these systems, we can infer why decomposition is considered an integral part of the new-mechanistic approach to explaining their behavior. I begin by presenting the concepts of composition and decomposition, examining how these complementary processes relate to revealing the hierarchical organization of biological structures across multiple levels (2.1. *From Parts to Whole and Back Again: Exploring Composition and Decomposition*). This discussion sets the stage for a deeper investigation into the nature of complexity in biological systems (2.2. *Breaking Down the Complex: Decomposition in Biological Systems*), where after defining and delineating the subject (2.2.1.), I explore various approaches to decomposing these systems and the inherent challenges involved (2.2.2.). I then focus on two key aspects of biological organization (2.3. *Structural Principles and Organizational Dynamics in Complex Biological Systems*). First, I explore the emergence of hierarchical structures through aggregation and sub-aggregation, drawing on Wimsatt's work to illustrate how restricted forms of aggregativity contribute to the formation of nested hierarchies (2.3.1.). Second, I delve into the role of interactivity and modularity in shaping biological architectures (2.3.2.). This discussion incorporates examples from diverse biological contexts, from molecular interactions to organ systems, to illustrate these principles in action. In the last part of the chapter, dedicated to the concept of near-decomposability (2.4. *Idealization and Near-Decomposability: Bridging Simplicity and Complexity*), I explore how this idea allows for idealized analysis of complex systems while still capturing both modular and emergent whole system properties. Throughout the chapter, I draw on a rich array of examples from various biological systems, ranging from cellular structures to ecosystem dynamics. These examples illustrate the practical application of decomposition techniques and theoretical concepts in understanding living organisms.

2.1. FROM PARTS TO WHOLE AND BACK AGAIN: EXPLORING COMPOSITION AND DECOMPOSITION

Here, I begin with a general definition of the concepts, *composition*, and *decomposition*. Firstly, the term decomposition appears in the science of complex systems and refers to the analytical process of breaking down a system and its behavior into smaller parts; it is an explanatory-oriented reconstruction where a complex whole is divided into simpler constituents. These could be material parts but also other, more abstract entities like functions/operations. Decomposition serves several purposes: simplifying a problem by breaking it into smaller, more manageable units, understanding the consistency of a system, identifying different aspects of a complex whole, unraveling the separate contributions and roles of its constituents, etc. On the other hand, composition³ refers to the synthetic counterpart of the decompositional process, involving the combination of many parts to form a whole at a higher level. This concept regards part-whole relations; if we have a system and several entities, it answers the question, are these entities integral parts of the system? Additionally, it concerns the organizational principles applied to multiple entities, resulting in their grouping, often across multiple levels of organization, to produce organized structures that constitute the whole. While complementary, these concepts describe processes with opposite directions: composition moves towards upper levels, while decomposition moves towards lower levels.

To illustrate, let's consider the *decomposition* of a system into concrete (material) parts. For example, to determine the types and quantities of constituents comprising the human body, one could use an analysis like the *five-level model of body composition* (Duda et al., 2019). This model identifies levels such as atomic, molecular, cellular, tissue-organ, and whole body, describing the components and their quantities at each level: At the atomic level, body mass consists mainly of oxygen (61%), carbon (23%), hydrogen (10%), nitrogen (2.6%), calcium (1.4%), and phosphorus

³ In many cases that are relevant to the discussion of explaining complex biological systems, the use of the term 'constitution' generally serves the same role as the term 'composition'. However, there are some variations in its use, mainly between Carl Craver (2007) and Carl Gillett (2016). Craver, in the context of his mechanistic framework, uses the term constitution to signify that an upper-level mechanism for a given phenomenon is made up, viz, is constituted by the contribution and interaction of entities and activities at a lower level. Gillett, on the other hand, uses the term 'composition' broadly to declare that many entities, which can be individuals, properties, powers, or processes, compose a whole at an upper level. Furthermore, he uses the term 'constitution' more narrowly to declare a special kind of compositional relation that applies only to individuals and not to the other three kinds of entities.

(0.83%). At the molecular level, it comprises water (approximately 60%), anhydrous fat (20-30%), body proteins (14%), minerals (5%), and glycogen (1%). At the cellular level, it includes extracellular space (consisting of extracellular fluid and solids, 8% of body mass) and cellular mass (36-46%, varying with factors like gender and health condition). Similar analyses can be applied to the remaining levels. In this model, each level corresponds to the size scale of the entities which result from the decomposition procedure.

Furthermore, at each size level, different decompositions can be made depending on the desired outcome. For example, at the cellular level, a morphological decomposition would list all existing cell types and estimate their quantities. Another decomposition might consider the functional role of cells in a system or organ, classifying them accordingly; for instance, the immune system at the cellular level consists of leukocytes (white blood cells), which are of six types: polymorphonuclear neutrophils, eosinophils, basophils, monocytes, lymphocytes, and plasma cells. This decomposition relates to their immune role. If cell morphology were the classification criterion, the same cells could be divided into two categories based on their granular appearance. The first three types of leukocytes are granular, possessing multiple nuclei that give them a characteristic appearance and are called granulocytes (Guyton & Hall, 2006). All these decompositions answer the question: Which concrete entities compose a particular entity at a specific size scale according to a given criterion? It's worth noting that the levels themselves are not given but may result from the recursive process of decomposition.

Let's now consider the opposite case of (material) *composition*. Composition deals with *parts*, *parthood*, and their organization as a whole. Let's first see parthood. This notion applies to both material things and more abstract entities. For instance, the statement "the veins are parts of the circulatory system" designates a material parthood relation, while the fact that the words "veins" and "circulatory" are parts of this statement refers to a more abstract, linguistic parthood relation. Hence, the point here is to formulate criteria by which something is or is not part of a whole. Again, as with decomposition, the decision depends on the specific request. For example, a regional decomposition of the human body divides it into the head, neck, thoracic cavity, abdominal cavity, pelvic cavity, and upper and lower limbs. Another decomposition might distinguish between anatomical categories, e.g., internal organs (viscera), locomotor apparatus, nervous system, and

skin, while a third might assign body parts to systems⁴ according to their physiology. We can explore composition by asking questions about the lung and heart: Are they part of the thoracic cavity? Are they part of the viscera? Are they part of the circulatory (cardiovascular) system? For the lung, the parthood relation holds only for the first two questions, while for the heart, all questions have affirmative answers. These questions utilize criteria relevant to organ position, anatomical category, and functional role in an organ system, respectively. If we ask the same questions about the femoral artery, we will get negative answers for the first two and an affirmative answer only for the third, as it is part of the circulatory system but does not reside in the thoracic cavity, nor is it considered an internal organ.

The organization of the component parts complements and completes the concept of composition and answers the question, how do parts compose an organized complex whole? There are several principles for the organization of parts and functions and will be discussed in more detail later in this chapter. The general rationale is this: an organizational principle groups many components into one at a higher level of organization. The recursive application of this process creates hierarchies of levels that reflect a system's internal structure.

What if we are not talking about concrete but more abstract entities, e.g., functions/operations or capacities? When dealing with abstract entities, both decomposition and composition concepts still apply; in mathematics, for instance, decompositional methodologies are widely used for solving complicated equations, as in the case of polynomial factoring. Functional decomposition also serves as an analyzing tool for computer science. In *Information Processing Theory*, one of its classic applications is the *flow diagram*, which depicts how a higher-level informational event is decomposed into a set of lower-level operations. Arrows designate the operations' interrelations and flow dynamics, such as the temporal organization of lower-level operations. By recursively applying this procedure, a *tree graph* is produced, representing the hierarchical structure of the flow diagram's components.

⁴ The functional subdivisions of the body organ system distinguishes four categories: locomotor system, comprised by skeleton, joints and musculature; viscera, comprised by cardiovascular, hemolymphatic, endocrine, respiratory, digestive, urinary and reproductive systems; nervous system which further divides in central, peripheral and sensory organs; the skin and its appendages (Schünke, Schulte, Schumacher, Ross, & Lamperti, 2006).

In the biological domain, functional decomposition is a typical process for investigating complex biological mechanisms. It was shaped mainly by Herbert Simon's (1962) pioneering influence on complex hierarchical systems, William Wimsatt's (1972) extended research on complex biological systems, and Robert Cummin's (1975) functional analysis schema. The latter's general idea is to explain a specific (analyzed) capacity through the organized cooperation of a set of (analyzing) sub-capacities. This process requires, firstly, a definition of the function to be explained (the analyzed capacity) and, secondly, a decomposition into an organized set of simpler functions (analyzing sub-capacities). The procedure continues recursively until it ultimately appeals to natural laws and principles. In a functional decomposition in the biological domain, a system's activity is considered the product of a set of subordinate functions or operations performed by several subsystems that interact to varying degrees.

Let's consider an example from the urinary system, specifically the kidneys, simplified for conceptual elucidation. Their function is to remove waste materials from the body and control the volume and composition of body fluids. We can decompose kidney function into two sub-functions (Guyton & Hall, 2006): urine formation and urine excretion. Urine formation can be further decomposed into three renal processes: (1) glomerular filtration, (2) reabsorption of substances from the renal tubules into the blood, and (3) secretion of substances from the blood into the renal tubules. We can continue this functional decomposition by analyzing glomerular filtration, for instance. Here, we will need to consider the interaction of forces that cause filtration by the glomerular capillaries, namely the glomerular hydrostatic pressure, the Bowman's capsule pressure, and the glomerular oncotic pressure. Taking these forces into account, we can estimate the net filtration pressure⁵, which determines the glomerular filtration rate, which in turn determines the volume of urine. In this example, three successive functional decompositions were applied to operations in a hierarchical manner. Each level resulted from the decomposition of the operation of the level above it.

The assessment of all the individual functions and all the individual components at each level provides varying degrees of detail about the overall functioning and structure of the system. The decompositional procedure may concern only two levels, the higher, which corresponds to the

⁵ The net filtration pressure represents the sum of the hydrostatic and colloid osmotic forces that either favor or oppose filtration across the glomerular capillaries. (Guyton & Hall, 2006, p. 337)

whole, and the lower, which corresponds to the parts. The recursive application of it results in a hierarchical system with multiple levels where any lower are nested on the higher ones. In studying a complex biological system, functional and material decomposition are often performed simultaneously and complementarily. This means that one not only identifies a particular set of processes and sub-processes that compose an upper-level function but also ascribes these processes to specific body structures that execute them. Complementarity also applies to decomposition and composition procedures, and any explanatory attempts oscillate between top-down decompositional analyses and bottom-up compositional reconstructions.

The examination of part-whole (compositional) relations enables identifying the system's boundaries, e.g., what belongs or not to it. But decomposition alone is not sufficient to explain a phenomenon; it needs to be accompanied by an account of the temporal, spatial, and active organization of the components and functions that comprise a system. In what follows, I will proceed to a more detailed presentation of decomposition as it is implemented in the analytic methodologies of complexity science. In the last part, I will refer to the works of Simon, Ando, Courtois, and Wimsatt about composing and decomposing complex biological systems.

2.2. BREAKING DOWN THE COMPLEX: DECOMPOSITION IN BIOLOGICAL SYSTEMS

In the preceding section, I presented the notion of decomposition quite simplified; the aim was to give a sketchy outline. However, since we deal with explanations in biological systems, which are paradigmatically complex, things are much more complicated. In this part, after a brief discussion about complexity in general and the kind of complexity involved in biological systems in particular, I will examine how a decompositional methodology deals with this specific kind of complexity. More generally, within the complexity science context, decomposition refers to many different types of analytic methodologies depending on the kind of complexity we deal with. In any case, a common denominator is that decomposition is one way of resolving the complexity of a system.

2.2.1. Understanding Complexity in Biological Systems

Complexity science's research subjects are complex systems. According to a standard definition, such systems consist of interdependent, diverse entities, responsive to their local and global environments, hence adaptive, that connect or interact with each other in various ways forming networks (Page, 2009). Sporns (2007), describes the system's entities or components as

structurally differentiated units that, through their individual functions and interactions, create some form of local behavior, activity, or dynamics. Entities themselves can be composed of other components, forming hierarchical systems that are complex at many organizational levels. Often, these hierarchies cannot be fully decomposed. Furthermore, the dynamic interactions of components in space and time integrate them into organized wholes. Interactions often regulate the components' individual actions, thereby changing their local functionality depending on the context in which they take place. The components have varying degrees of interaction, strong or weak (Sporns, 2007). The entities that comprise a complex system have the ability to self-organize into the spontaneous formation of temporal, spatial, or functional structures, and very often, the collective behavior of the whole system cannot merely be inferred from the understanding of the behavior of its individual components, meaning that those systems give rise to emergent behaviors and phenomena (Meyers, 2009).

Although complexity science aims to study, analyze, and explain complex systems and their phenomena, there is no identifiable global complexity theory that can serve this general purpose; instead, there are many different theories and methodologies related to several diverse disciplines, perspectives, and types of phenomena. According to Manson (2001), it is possible, however, to distinguish three major divisions of complexity research:

“Algorithmic complexity,” in the form of *mathematical complexity theory* and *information theory*, contends that the complexity of a system lies in the difficulty faced in describing system characteristics. “Deterministic complexity” deals with *chaos theory* and *catastrophe theory*, which posit that the interaction of two or three key variables can create largely stable systems prone to sudden discontinuities. “Aggregate complexity” concerns how individual elements work in concert to create systems with complex behavior.” (Manson, 2001, p. 405)

In this last case, researchers study systems comprised of many components, aiming to access the holism and synergy resulting from their interactions. In these systems, the emphasis is more on the relations than on the constituent parts. The relations define the system's functions and goals, resulting not from a centralized design; instead, they emerge from the components' interactions. The relations between the components vary; some have notably tight connections and form subsystems, although any given component can belong to multiple subsystems. The former characteristics relate more closely to the kind of complexity involved in biological systems, given

that the relevant entities, e.g., cells, organisms, and ecosystems, are complex because they consist of a large number of mutually interacting parts.

An especially illustrative example of a quasi-biological system characterized by aggregate complexity comes from studying emergent behavior using cellular automata models. In a recent study, Schmickl, Stefanec, and Crailsheim (2016) developed a model for the spontaneous development of life-like structures that emerged through the interactions of particles that obeyed a simple motion law. A population of particles moves in a deterministic and asynchronous manner in a grid that corresponds to continuous space, and each particle is defined by its position, heading, and movement. Furthermore, within a certain radius, it interacts with other particles resulting in a change of its heading, depending on the number of other particles to their left and right. At a specific parameter set and after certain time steps, the particles start to aggregate, and this configuration results in the spontaneous emergence of many impressive self-organized structures and behaviors. The simple kinetic behavior of the particles at the lower level of description gives rise to cell and spore-like structures at a second, higher level, and at a third, even higher level, one might observe an eco-system-like organization with emergent life cycles where the structures grow, reproduce and interact with each other. Clearly, the *ecosystem* and the *organisms* that emerge there are far from real biological organizations, but this research provides an essential proof of concept for how self-organized structures emerge from systems of aggregate complexity and offers useful insights into how to study and explain them. Many parts, simple behaviors, and a sufficient degree of interactivity at a lower level may result in profoundly complex systems at a higher level.

Let us now proceed to a more thorough exploration of the term *biological complexity* as it is defined and studied within the Complex Systems Approach discipline. A widely accepted definition, proposed by one of the most prominent complexity theorists, Chris Adami, is the following: “Biological complexity refers to a measure of the intricateness, or complication, of a biological organism that is directly related to that organism's ability to successfully function in a complex environment” (Adami, 2009, p. 490). What exactly does that mean? For instance, what makes an organism more complex? It is generally accepted that a virus or a bacterium is more likely a simpler organism than a mammal and, in particular, humans. An assessment like that depends on criteria such as the large number and variety of cells or tissues that compose a human organism combined with the functions and interactions that take place at each hierarchical level of organismic description. However, it seems that an equally important role plays the complexity of the environmental niche where an organism struggles to survive and prosper, in other words, to adapt.

For example, the human being, commonly regarded as the *crown jewel* of evolution, can cope with a wide variety of environmental conditions and different worlds of existence, such as the physical, the mental, the social, et cetera. Hence, internal and external factors orchestrate biological complexity.

Adami (2009) reviews several proposed ways of evaluating and measuring biological complexity. Regarding their applicability to the various complexity research divisions, they can be divided into two groups of measuring methods; the first deals with algorithmic complexity and utilizes information-theoretic and mathematical analyzing methodologies, while the second relates more to the aggregate complexity approach. The first group of methodologies aims to measure complexity and informational content in macromolecules, especially in DNA, by evaluating their sequence complexity. For instance, regarding the genetic code, researchers measure the sequence length and apply mathematical tools to estimate several types of sequence complexity like the Kolmogorov⁶, compositional⁷, and information-theoretic⁸ complexities. The second group of methodologies relates more to aggregate complexity and focuses on an organism's *structural, functional, and organizational* aspects.

According to Adami (2009), evaluating a system's complexity can be accomplished through specific compositional analyses focusing on different aspects of the system. For instance, assessing its structure, based mainly on the number and diversity of parts and their connections (e.g., the number of different cell types within an organism), is likely a relatively crude estimate of complexity. He notes that a more precise calculation of structural complexity also evaluates an organism's hierarchical structure, i.e., the number of levels comprising a specific biological system's hierarchical organization. Hierarchical complexity evaluation proceeds with a recursive

⁶ *Kolmogorov Complexity*, named after the Russian mathematician Andrey Kolmogorov, who introduced this measure, aims to evaluate the regularity and, consequently, the descriptiveness of a sequence to determine its complexity. A sequence that is difficult to describe computationally is considered more complex. For example, a string without any discernible pattern is irregular and thus more complex. Conversely, a regular string will have easily identifiable symbols or patterns that repeat, resulting in a shorter description. In essence, the more concisely a string can be described, the lower its Kolmogorov Complexity. (Adami, 2009).

⁷ *Compositional complexity* aims to capture sequence complexity by evaluating its heterogeneity. To achieve this, a criterion is applied to find the optimal segmentation of the sequence into a number of partitions. A highly heterogeneous sequence is characterized by high compositional complexity (Adami, 2009).

⁸ *Information-theoretic* approaches deal with biochemical information, with the measure of sequence entropy playing a leading role in the early applications of information theory. Later, however, the focus shifted to measuring informational content, and many sophisticated methods were developed to unravel the depth of biochemical informational content in macromolecules and genes (for further information, see Adami, 2009).

decomposition of an organism into entities residing at many levels and sub-levels, introducing a higher-resolution scale. Additionally, he proposes to measure the functional complexity of a system, which reflects an organism's behavior in a complex environment. A quantitative estimation based on the diversity of functions that constitute an organism's performance is perhaps impractical, as the organism's functions are co-defined by environmental conditions, introducing a high degree of uncertainty. An indirect approach would be to deduce these functions by identifying the organism's different observable behaviors. This could be done recursively by decomposing behaviors/functions at every level and for each working part.

Adami (2009) stresses that a complete evaluation of the aggregate organismic complexity should include measuring the organization of functions and material components, representing the network's complexity. An organism can be studied as a functional network, and biological structures are characterized by organizational patterns that are far more complex than several nodes randomly connected. One of the most typical features found in biological networks is the property of modularity, i.e., the functional specialization of specific parts of the network that in the network graph has the appearance of a cluster consisting of more intensively connected nodes. Often, several nodes can belong to more than one module at a time, which increases the system's complexity and integration. In any case, the way to evaluate the biological network is to compare it with a random graph and make a qualitative assessment of the modularity that exists in it. Several other methodologies have been developed within Graph Theory for analyzing networks and extracting information through them⁹. All these strategies contribute to revealing the organization of a complex biological system. Hence, structural and functional decomposition and depicting a system's organization constitute the three main ways of measuring aggregate complexity in an organism.

2.2.2. Approaches to Decomposing Biological Systems

Before proceeding in more detail to decompose structurally and functionally biological organisms, one must keep in mind something vital: every decomposition, whether it concerns structures or functions and interactions, is based on some criterion. In other words, there is not only one decomposition but many, and it depends on what we want to find. This has been particularly emphasized by Stuart Kaufmann (1970). He distinguishes two diverse targets of scientific

⁹ For other methodologies used to resolve the complexity of a network, consider Adami (2009).

explanation: firstly, things that are studied as if they are a single part or a continuum, and secondly, things that are constituted of parts and are explained as an articulation of parts. In the first category, an explanation for a specific behavior does not need to exhibit how parts work together. Note, however, that a single system can exhibit several different behaviors, and some of them may be explained as referred to a continuum while others may require an account of how parts work together. We can refer to these two categories that Kaufmann suggests as *non-decomposable* and *decomposable* behaviors.

He proceeds to the second category and notes that decomposition into parts and processes depends on the behaviors (or phenomena) that someone wants to study. Studied behaviors and decompositions are mutually dependent. Therefore, the procedure is bidirectional: we use various views of what a system does to guide its decomposition into parts, and accordingly, we use information about parts to synthesize new views of what a system is doing. Consequently, Kaufmann suggests that there may be many different decompositions for a given system according to different behavior descriptions. When two descriptive decompositions are applied, it is sufficient just to be compatible with each other. There is no need for an ultimate decomposition from which all the others are deducible. So, we first need to get an initial idea of the system's behavior. This provides the explanandum and sets criteria for whether a particular section of the system and several of its causal consequences can be considered parts and are involved in the system's processes. In practice, someone evaluates jointly different parts and processes, given that they work together to produce the behavior under investigation. Kauffman's ideas reflect especially one of the key features of aggregate complexity, namely that the entities that make up a system, whether they are physical parts or functions, interact with each other; thus, they are co-determined.

Wimsatt (1972) further develops Kaufmann's considerations to a more extensive account of decomposing complex systems. He distinguishes two types of complexity: *descriptive* and *interactional* complexity. In the first type, we describe a system, e.g., we identify its boundaries and constituents, utilizing several criteria based on theoretical perspectives, according to which we can decompose it in parts. Every different criterion corresponds to a different decomposition and, accordingly, to a different description of parts. The results of these decompositions reflect the system's descriptive complexity, depending on whether there is a spatial coincidence in the parts' boundaries resulting from the decompositions. Wimsatt puts it in a formal way: if we suppose T_i different theoretical perspectives, they imply different criteria for identifying and individuating the system's parts, which produce $K(T)_i$ decompositions. These decompositions may or may not result

in spatially coincident boundaries for some or all the parts of the system, and accordingly, the descriptive complexity of the system is evaluated:

If all of a set of decompositions, $K(T)_i$, of a system, produce coincident boundaries for all parts of the system, the system will be called *descriptively simple* relative to those $K(T)_i$.

If two parts from different $K(T)_i$ are not coincident but have a common point, which is an interior point of at least one of them, then there are a number of different mapping relations which can hold between their boundaries, each of which contributes to its *descriptive complexity*. Specifying these mapping relations for all parts of the system under both decompositions gives a complete description of this complexity from a set-theoretic point of view. (Wimsatt, 1972, p. 70)

Wimsatt (1972) uses examples of a granite piece as a descriptively simple system (relative to a specific set of $K(T)_i$ decompositions) and a multicellular organism as a complex system. The granite can be decomposed based on criteria such as T_1 (chemical composition), T_2 (density), and T_3 (thermal conductivity). The resulting decompositions, $K(T)_1$, $K(T)_2$, and $K(T)_3$, will provide subregions with approximately coincident boundaries. In contrast, a multicellular organism can be analyzed using T'_1 (physiological), T'_2 (anatomical), and T'_3 (biochemical) criteria for $K(T')_1$, $K(T')_2$, and $K(T')_3$ decompositions. The resulting different cell types, anatomical organs, and biochemical reaction types correspond to neither isomorphic nor spatially coincident mappings. Hence, this is a descriptively complex system relative to the specific decompositions.

Wimsatt also refers to another type of complexity that attempts to capture in great depth the inherent complexity of a system and the difficulty of formulating theories of its operation that will provide accurate predictions. This second type relates to the interactionist profile of a system¹⁰. He calls it *interactional complexity*, and a prerequisite for understanding it is the concept of *near-complete decomposability* that was first introduced by Simon and Ando (1961). Let us first briefly examine this notion. A system can be decomposed into subsystems taking into account the strength of the causal interactions between its constituent parts. In this case, subsystems are defined

¹⁰ It is worth noting that in the rapidly growing field of Complex Systems Approach, the primary focus is on the interconnections and relations within a system rather than its individual components (Aon & Cortassa, 2009). This emphasis appears to stem from the fact that these interconnections and relations are responsible for the most distinctive properties exhibited by complex systems, such as self-organization, adaptation, emergence, and evolution over time.

dynamically, irrespective of spatial or morphological criteria. In a system, some parts interact strongly, some weakly, and others not at all. Those that interact strongly are clustered in subsystems, and all the strong interactions are contained within them. Since there may be several weak interactions between parts registered in different subsystems, the system is characterized as nearly decomposable, and the accuracy of the resulting subsystems depends on how much stronger the strong interactions are compared to the weak ones. The parameter $\mathcal{E}_c$, which is the ratio of the strongest weak force to the weakest strong force, reflects the maximum error produced if weak forces are ignored, and it measures the degree to which a system is nearly or completely decomposable.

Suppose, now that we use a T_1 descriptive criterion to produce a set of parts, then measure the interactions and divide them into a set of subsystems. Let us further assume we use another T_2 descriptive criterion, which also decomposes our system into a set of parts. And similarly, we evaluate causal interactions for the corresponding parts to break it down into subsystems. If the subsystems resulting from the two near-decompositions do not differ significantly from each other, that is, if no strong interactions occur between parts of the one decomposition that were weak in the other decomposition, then the system is interactionally simple. However, if strong interactions occur that traverse between the two different descriptive decompositions, then the system is characterized as interactively complex. In Wimsatt's (1972, p. 72) words, "this is a kind of measure of the complexity of the causal interactions of a system, with special attention paid to those interactions which cross boundaries between one theoretical perspective and another." If a system is simultaneously descriptively and interactively complex, we need to consider more than one theoretical perspective to make accurate predictions for its behaviors. Consequently, how complex a system is is judged at the descriptive level, at the level of interactions, and the combination of the two above.

To recapitalize the whole schema as it arises from what has already been presented, biological organisms are characterized by aggregate complexity, which involves systems composed of numerous parts and functions related to allowing the system to adapt to a complex environment. These systems can be broken down into dynamically organized structures of component parts and component functions/operations. They are always decomposed according to specific criteria related to system properties or phenomena under examination. Initially, a descriptive decomposition into entities by selected functional or morphological criteria put the bases, to be followed by the investigation of the relationships that entities have with one another, that is, the

interactional decomposition of the system. Thus, the system can be divided into subsystems, and its hierarchical structure is revealed. Such a system's complexity is due both to the diversity and number of entities that constitute it and the interactions between them.

2.3. STRUCTURAL PRINCIPLES AND ORGANIZATIONAL DYNAMICS IN COMPLEX BIOLOGICAL SYSTEMS

Three key elements dominated the preceding discussion about the explanatory decomposition of a complex system: material parts, functions (or operations), and interactions. They are interrelated, and, to some extent, there is some difficulty in clearly distinguishing between operations and interactions. Breaking down a system into its component parts primarily offers insights into its physical composition. However, for a biological organism, which is a living entity with observable behavior, explaining this behavior necessitates decomposing it into distinct operations that collectively give rise to it. Decomposition into component parts and operations is a prerequisite for examining interactions because one must first identify certain parts before investigating whether, how much, and how they interact. There is, though, a difficulty in distinguishing between the notions of *functions/operations* and *interactions* since each operation is also an interaction. In essence, something acts (or operates) on something else, i.e., interaction is an operation, but the interactions also include clues for a system's functional organization and provide insights as to how this can be broken down into subsystems that perform subfunctions. Nevertheless, an explanatory strategy can focus on investigating only one of the two elements: either studying only the interactivity between a set of components and identifying the formation of functional subsystems among their interactions or else decomposing a function into subfunctions, ignoring their interactions and patterns of organization.

Bechtel and Richardson (2010) discuss explanatory attempts in presumably hierarchical and decomposable systems and phenomena, focusing on detecting component parts and operations through two complementary strategies. The first is decomposition, a top-down strategy where a phenomenon or function is decomposed into component operations (subfunctions). The second bottom-up strategy is localization, where components are identified, and their roles in performing

operations are examined¹¹. Additionally, they note that a functional decomposition initially does not involve simultaneously investigating operations and interactions. At the beginning of the analysis, operations are considered minimally interactive and mostly linear. Although unrealistic, this assumption is necessary for the preliminary analysis to be performed. In other words, we start with the assumption that a system is actually decomposable, which means we assume it has a relatively simple, modular organization.

In the complex phenomena research field, the distinction between components versus interactions is notably essential, emphasizing interactions, which are vital in Complexity. If a system can be decomposed into component parts or component functions, it is governed by the so-called *component-dominant dynamics* - ideally, it is completely decomposable. In contrast, in systems where interactions are numerous, nonlinear, and complex, the system is governed by *interaction-dominant dynamics* and, in many cases, is characterized as non-decomposable (Silberstein & Chemero, 2011). This distinction has been widely used in the *Dynamical Approach to Cognitive Science* context as well. According to Silberstein and Chemero (2011, p. 4): "In component-dominant dynamics, behavior is the product of a rigidly delineated architecture of components, each with pre-determined functions; in interaction-dominant dynamics, on the other hand, coordinated processes alter one another's dynamics, with complex interactions throughout the system." These interactions are strongly nonlinear and can extend to various spatial or temporal scales resulting in the inability "to determine the contributions of the individual working parts to the overall operation of the system; in a system whose dynamics are interaction dominant, all of the system's parts are constitutive" (p. 5).

In any case, to analyze a system in terms of its component parts, functions, interactions, and overall organization, a degree of idealization of either components or interactions is necessary. When investigating patterns of interactions, the diversity of components is often simplified or overlooked. This simplification, however, is far from reality, as the heterogeneity of a system's components significantly contributes to its ability to produce emergent behaviors - a hallmark of complex systems, which biological organisms are widely considered to be. Similarly, when focusing on

¹¹ The distinction of (functional) decomposition vs localization is elsewhere referred by Bechtel (2008) as two complementary procedures of functional and structural decomposition, with the first corresponding to the decomposition of a system into working parts which are involved in operations. Those operations are detected through the functional decomposition of a system. Localization is the assignment of an operation to a certain part.

components, an idealization assumes that a system is completely decomposable, implying that interactions are linear, simple, and well-defined within specific subsystems. In the following sections, in order to gain a deeper understanding of complex organized wholes, I will explore two factors that concern how structure emerges in biological systems and determine their decomposability, thereby influencing the extent to which they can be studied by explanatory strategies that presuppose decomposition. First, I will delve into the role of material components and their organization in shaping level hierarchies that correspond to systems exhibiting aggregative or emergent properties. Subsequently, I will consider the role of interactions in the formation of a modular structure and the idealized situation of a completely decomposable system. This will lead us to the last section of the present chapter, where the discussion will revolve around the extent to which we can apply a degree of idealization to a nearly decomposable system to explain it as if it were fully decomposable.

2.3.1. Aggregativity vs Emergence: When Parts Become More Than Their Sum

Wimsatt (1986), in “Forms of Aggregativity,” conducted a thorough inquiry into what makes systems mere *aggregates* of their parts and how they differ from systems that are *complex organized wholes*. Several properties may characterize a system as a whole, and each analysis targets one of these. The question is whether a given property is emergent or aggregative. That is, a system may have properties of an aggregative character while other properties are less aggregative or more emergent.

For example, consider a pile of stones where the properties under examination are the weight and height of the pile. The weight is more aggregative than the height since rearranging the stones in space will not change the pile's weight, while the height may change. In the latter case, the spatial organization of the components plays a vital role, but this is irrelevant to the weight. It's essential to keep in mind that emergence and aggregativity are the two distal points of a continuum, and emergence can be seen as a failure of aggregativity. Any property of a system could be positioned somewhere between these two extremes.

To examine Wimsatt's ideas, let's consider the internal structure of a system, which, according to a specific decomposition, consists of numerous component parts with their properties, arranged in space and time and interacting based on several organizational principles. Let S be the system and $s_1, s_2, \dots, s_m$ be the system's components, each having n properties: $p_1(s_1), p_2(s_1), \dots, p_n(s_1), p_1(s_2), p_2(s_2), \dots, p_n(s_2), \dots, p_1(s_m), p_2(s_m), \dots, p_n(s_m)$. The system's property or phenomenon under

examination, $P_j(S)$, is determined by the spatiotemporal arrangement and interactions of its parts and is aggregate insofar as four conditions are met. The first three conditions primarily concern the descriptive complexity of a system, while the fourth addresses its interactional complexity. The following is the *composition function* of a system's property $P_j(S)$, where F represents the theory that associates all the components and their properties with the property or phenomenon that the whole system exhibits (Wimsatt, 1986, p. 260):

$$P_j(S) = F [p_1, \dots, p_n(s_1), \dots, p_1, \dots, p_n(s_m)]$$

The four conditions that Wimsatt (1986) proposes are the following:

Condition IS: $P_j(S)$ is invariant under intersubstitution of the parts of S with one another or under substitution of one or more of the parts with other parts from a domain of relevantly similar parts. (p. 262)

Condition QS: The system property remains qualitatively similar (differing only in value) under addition or subtraction of parts. (p. 264)

Condition RA: The composition function for $P_j(S)$ is invariant under operations involving decomposition or reaggregation of parts. (p. 266)

Condition CI: There are no cooperative or inhibitory interactions among the parts of the system. (p. 269)

If all four of Wimsatt's proposed conditions are met for a system property, this property is considered aggregative. Any deviation from these requirements results in properties with a more emergent character.

Firstly, we will examine the role of the organization of a system's parts. Consider a system S with all its components having the same property (p_1), such as weight. Using the pile of stones example, the total weight $P_1(S)$ equals the sum of all stones' weights: $P_1(S) = p_1(s_1) + p_1(s_2) + \dots + p_1(s_n)$. This remains constant even if we rearrange their positions or replace one stone with another of the same weight. Thus, the weight property is aggregative for this system because $P_1(S)$ is "invariant under intersubstitution" (Wimsatt, 1986, p. 262) (Condition IS). The former is the first condition of aggregativity. However, if we calculate the height $P_2(S)$ of the pile, replacing a stone with a similar one in a different position would change the height, as it's not simply the sum of individual stone heights. Other factors, like each stone's position, play a role. This property doesn't meet condition IS because it depends on both the parts and their organization.

Moving on to the stone pile's scenario, if we remove a few stones, the pile's weight will change, but this will only be a quantitative change, and there will be no qualitative difference. The same stands for the pile's height: it would change as a value but not qualitatively. Both properties remain “qualitatively similar under addition or subtraction of parts” (Wimsatt, 1986, p. 264) (Condition QS). But if the property we are interested in is the stability $P_3(S)$ of the pile and the stones we removed supported some other stones, the pile could collapse. Intersubstitution can be made as far as we only exchange positions between two morphologically similar stones. Still, we cannot proceed to addition and subtraction without letting a significant qualitative difference occur. Thus, the property $P_1(S)$ (weight) obeys both conditions, the property $P_2(S)$ (height) obeys only the second condition, and the property $P_3(S)$ (stability of pile) obeys only the first. Therefore, if a system property does not follow one or both conditions, it is affected by how the system's parts are organized and, therefore, is not totally aggregate.

We can observe these concepts in a biological example concerning muscle force production within muscle fibers (i.e., muscle cells) through their length reduction¹². Each muscle fiber contains a large number of parallelly arranged myofibrils of the same length. Its ability to contract is due to the contraction of myofibrils. Two properties are considered here: the reduction in muscle fiber length during contraction and the force exerted on its ends. Regarding the first, due to the parallel arrangement of the myofibrils, the muscle fiber's total length reduction is the same as the individual myofibril's length reduction. Even if we replaced one myofibril with another of the same length, we would have the same length reduction. Consequently, the change in muscle fiber length resulting from the myofibrils' length change is an aggregative property because it is invariant under intersubstitution (condition IS). However, the second property behaves differently and is less aggregative: the total force produced at the end of the muscle fiber is equal to the algebraic sum of the individual forces produced in its myofibrils. If we remove or add another myofibril, the force will not change qualitatively but will change quantitatively. A thicker myofibril will not change the muscle fiber length but will make its contraction stronger. Therefore, the produced force's property is qualitatively similar under the addition or subtraction of parts, but it is not quantitatively the same (condition QS). In both cases, we are dealing with parts that are the same in some of their

¹² For more detailed descriptions of the relevant mechanisms consider Davis, Tikunova & Janssen (2019) and Oatis (2009)

properties and relatively simply organized, so the upper-level properties are more or less aggregative at this level of analysis.

To be considered aggregative, a property must satisfy more stringent criteria related to multiple descriptive decompositions of a system. These criteria specifically address how the system's components are organized and grouped into subsystems, as well as how the property behaves across various ways of dividing and recombining these parts. This requirement, which forms the third condition for aggregativity, ensures that the property remains consistent regardless of how the system is conceptually or physically partitioned; that is, a system is aggregative in terms of a property when it is invariant across different decompositions and reaggregations of parts. Here, let us recall the concept of Wimsatt's (1972) descriptive complexity and his relevant examples, presented in the previous section, and the characterization of some systems as *descriptively simple*, namely systems in which several different decompositions produce coincident boundaries for all the system's parts, as applies for example to a piece of granite. In such a system (e.g., the piece of granite), if we examine properties like the chemical composition or the whole system's thermal conductivity, they will remain the same whether we apply a decomposition based on the first or based on the second property. A consequence of this is that no matter how we divide it into subsystems and reunite them, these properties will not change because there is an overt uniformity in the system's structure. Such a system is of an aggregate character because it satisfies the third condition, which is: "the composition function for $P_i(S)$ is invariant under operations involving decomposition or reaggregation of parts" (Wimsatt, 1986, p. 266) (Condition *RA*). However, the same is not valid for the multicellular organism's example. In this case, several different decompositions will provide different sets of parts without spatially coincident boundaries and, consequently, with different properties.

Finally, in the fourth condition, we are dealing with any interactions between a system's parts. These can be cooperative or inhibitory, and their presence distances the system and the property under consideration from the aggregate end of the spectrum. That is, for a system's property to be aggregative, it must satisfy the condition that "there are no cooperative or inhibitory interactions among the parts of the system" (Wimsatt, 1986, p. 269) (Condition *CI*). In other words, the more interactions there are, the more likely it is that purely emergent phenomena, such as threshold phenomena, will occur.

Threshold phenomena occur when a property that depends on the mutual interaction of a system's components undergoes a dramatic change once a critical point is reached. This change is often characterized by a sudden, qualitative shift in the system's behavior. A well-known example of such a phenomenon in neuroscience is the initiation of a nerve action potential. In this process, the electrical properties of a neuron's membrane change abruptly when the membrane potential reaches a specific threshold, triggering a cascade of events that propagate the neural signal. This threshold-dependent behavior is influenced by the cooperative interactions among ion channels and other molecular components in the neuron's membrane. In this example, the property of the system as a whole is the permeability of sodium ions through the nerve cell membrane. After a gradual change of the resting membrane potential (-90 millivolts) to a less negative state, as soon as the membrane potential passes a certain threshold - usually between -70 and -50 millivolts - a sudden increase in the permeability of the sodium channel gate takes place. This results in a rapid entry of sodium ions inside the nerve cell, causing further depolarization of the membrane. This self-sustaining increase owes its existence to a positive feedback loop: the activation of voltage-dependent sodium ion conductance increases their entry into the neuron, resulting in further depolarization of the cell, a further increase in sodium ion conductance, and so on. This is the cooperative interaction part of the story. A parallel, slower negative feedback loop also occurs, which concerns the membrane's permeability to Potassium ions, which exit the cell, so eventually, the nerve cell membrane is restored to its resting state¹³.

In summary, a system's property is more aggregative when its parts are more uniform, less dependent on their organization, more constant across different decompositions, and less affected by cooperative and inhibitory interactions among the parts. Conversely, the above factors determine the degree of complexity of the system. It should be kept in mind, though, that being a fully aggregative system is an ideal case that is far from reality. First of all, the examples mentioned above considered specific whole-system properties that resulted from only one kind of property ascribed to multiple parts. However, especially in biological systems, the composition function that defines a particular system's property generally involves several different properties of the system's parts. This situation is called *complex aggregativity*. For instance, the sour taste

¹³ The mentioned mechanisms are detailed in (Purves, et al., 2004) "Neuroscience" Ch. 3: Voltage-Dependent Membrane Permeability and in (Guyton & Hall, 2006) "Textbook of Medical Physiology" Ch. 5: Membrane Potentials and Action Potentials.

impression caused by food consumption depends on the acidity impression caused by its ingredients. Namely, a single property (acidity) is involved that concerns many component parts (ingredients of food). However, if we consider the overall taste impression, which is more complex and depends on many other taste receptors, it will be determined by a combination of acidity, saltiness, sweetness, etc., caused by the ingredients and factors such as their textures, oiliness, and more generally, the quantitative and qualitative mixing of many properties related to the sense of taste.

Furthermore, in real life, we confront systems characterized by several kinds of restricted aggregativity, such as *local aggregativity*, where some or all of the conditions of aggregativity apply only to specific neighboring areas of the system. It is particularly important to note here that although a system that exhibits complete aggregativity may be biologically mundane, its complete absence may also be incompatible with biologically interesting phenomena. The presence of various forms of restricted aggregativity is an essential organizational principle for forming structures within a system, and this will be considered hereupon.

The case of restricted aggregativity, which is particularly interesting for the present inquiry, is *sub-aggregativity* because it is related to the organization of parts in a hierarchy of nested levels. Sub-aggregativity is the aggregativity within subsystems and is defined as: “limited intersubstitutability of components within subsystems, but none across subsystems” (Wimsatt, 1986, p. 271). This means that although the whole system may not be aggregative, the uniformity of some of their parts in at least one property makes them members of the same subsystem. The condition of intersubstitutability applies only within this subsystem. The concept of sub-aggregativity can also be applied to the subsystems. If a system's property is “sub-aggregative for some new (higher-level) subsets, (...), then the same operation can be applied again on the second-order subsets as parts” (Wimsatt, 1986, p. 275). Here lies the key to the formation of a hierarchical structure. The recursive application of these steps can generate a hierarchical structure. *Figure 1* illustrates the formation of a hierarchical structure through sub-aggregation within a collection of component parts, which are represented as circles and ellipses. The diagram shows multiple levels of organization, represented by circles and ellipses of varying sizes and colors. At the bottom level, numerous small circles in different colors (red, pink, green, orange, etc.) represent individual components. These are grouped into larger ellipses of corresponding colors at the next level up, demonstrating the principle of sub-aggregativity. This process continues upwards, with larger green ellipses encompassing groups of colored ellipses, ultimately leading to a single large green ellipse at the top. This visual

representation shows how components with similar properties (represented by colors) are grouped together at different levels, forming a hierarchical structure. The presence of gray circles throughout the lower levels represents components that are not part of the colored groupings, further emphasizing the selective nature of sub-aggregativity.

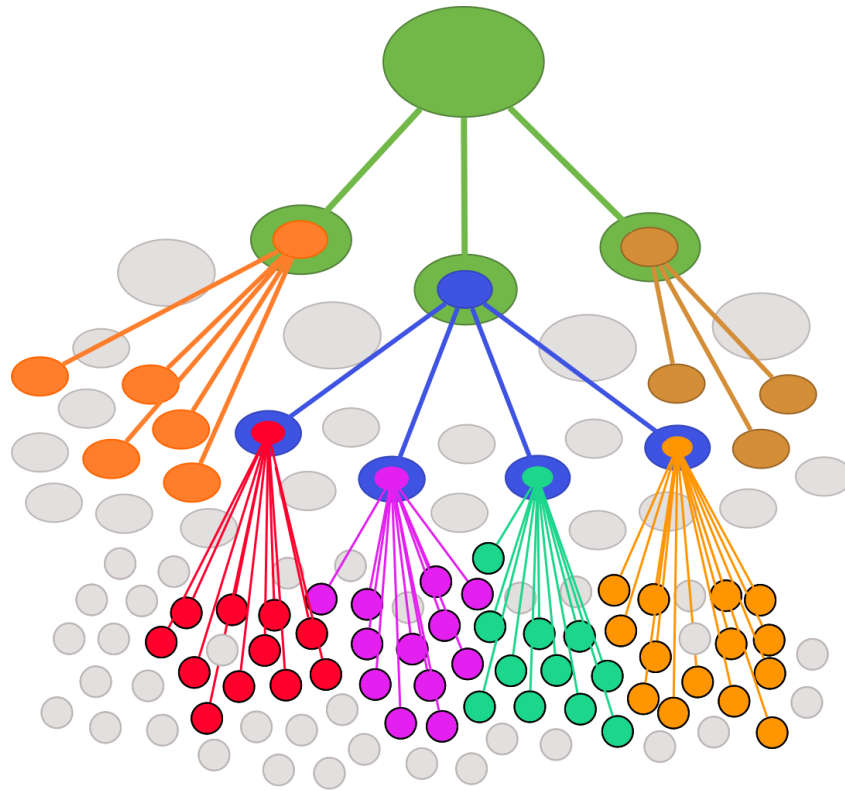


Figure 1: Formation of a hierarchical structure through sub-aggregation. The diagram illustrates multiple levels of organization, from individual components (small colored circles) at the bottom to a single encompassing structure (large green ellipse) at the top. Colored groupings represent sub-aggregative units with similar properties, demonstrating intersubstitutability limited only within subsystems. Gray circles represent components not part of the colored groupings, highlighting the selective nature of sub-aggregativity. This hierarchical structure emerges through the recursive application of sub-aggregation principles across different organizational levels.

The organizational principle of sub-aggregativity can be applied recursively to component parts and component functions that comprise functional units at higher levels. These higher-level functional units accomplish higher-level tasks, and so on, provided that in each higher-level sub-aggregativity can be applied. Numerous such examples can be found, for instance, in biology and computer science. In the biological domain, exemplary cases regard biological structures within individual organisms and groups of individuals within populations, with groups jointly taking on different functional roles within an ecosystem. An example of the application of sub-aggregativity that comes from the field of study of complex biological systems using Agent-Based Models is the concept of *super-agents* (Parry, 2009), as illustrated in *Figure 2*. These models try to explain and

predict phenomena, e.g., biological, which involve many active and interacting components, called *agents*. One option to simplify the calculations required in modeling such a large system “is to implement an aggregation of the individual agents (...) that share the same characteristics, such as age and sex” into super-agents (Parry, 2009, p. 150). This process can be repeated recursively on super-agents to encode a large amount of information through groupings that take advantage of the principle of sub-aggregativity.

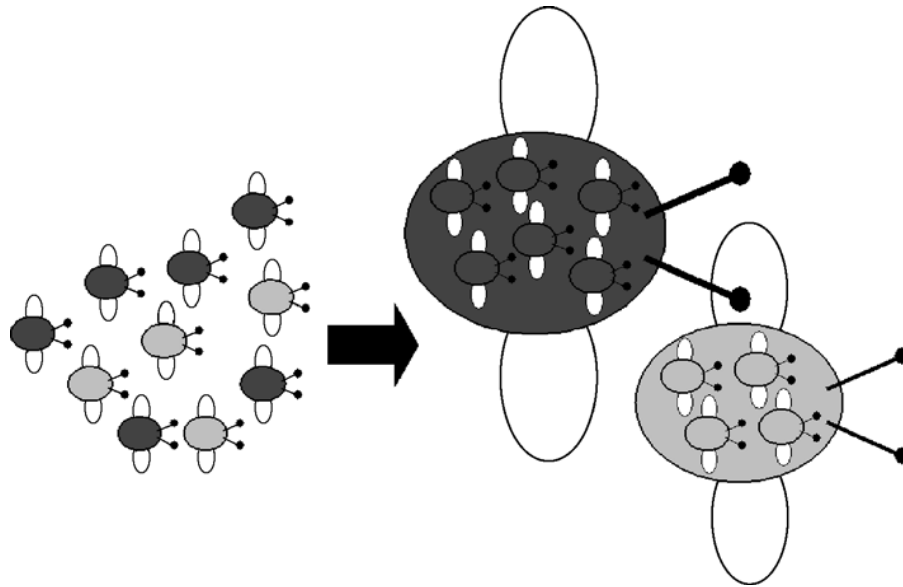


Figure 2: 'Super-agents': Grouping of individuals into single objects that represent the collective. Reproduced from Parry (2009, p. 151)

In conclusion, according to Wimsatt (1986), there are several ways to produce hierarchical organization, and restricted aggregativity is one. Another way is the identification of subsystems through the interactions of their components. In this case, subsystems are defined dynamically and relate to specific functions (behaviors). The emergence of structure in a complex system based on the interactions of its parts will be presented in the following section. Herbert Simon, whose ideas were inspired by the General Systems Approach, was the pioneering figure in the related ideas that concern their architecture as complex systems and their characterization as nearly decomposable, which has severe implications for their behavior and study.

2.3.2. Interactivity and Modular Organization in Biological Structures

Biological systems are living organisms or their functional parts, viz., organs. Life is characterized by properties such as persistence through time, adaptation, evolution, and reproduction. These properties require numerous structures that are functionally differentiated to accomplish various

goals. Analyzing these systems always reveals multiple interactions between entities at many levels of organization. Two crucial and interdependent concepts we will examine are interactivity and modularity. Interactivity refers to the dynamic exchanges and mutual influences between components within a system, which shape the system's behavior and properties. Modularity, on the other hand, refers to the formation of functionally specialized and autonomous structures within a system called modules. These concepts are fundamental to understanding the complexity and organization of biological systems.

The existence of interactions is key to understanding complex systems within the framework of the General Systems Approach. This approach, developed in the mid-20th century, identifies common properties, principles, and aspects among systems belonging to different fields, such as physical, biological, or social systems. It attempts to investigate their behavior from a synthetic rather than analytical perspective, focusing on the whole system rather than its individual parts. Ludwig von Bertalanffy, a pioneering biologist and the inspiration behind General Systems Theory, explains: "Its subject matter is the formulation of principles that are valid for 'systems' in general, whatever the nature of their component elements and the relations or 'forces' between them. General system theory, therefore, is a general science of 'wholeness' (...)" (Bertalanffy, 1969, p. 37). The most crucial postulation of this approach, essential for our discussion, is its emphasis on the interactions between the elements that make up a system. According to Bertalanffy, this feature is responsible for the characteristic assertion that "the whole is more than the sum of its parts" (p. 55). This concept, known as emergence¹⁴, reflects the fact that a system's behavior cannot be understood solely by studying its constituent elements. It is also necessary to examine the interactions between them, as these relations change the behavior of the parts:

"A system can be defined as a complex of interacting elements. Interaction means that elements, p , stand in relations, R , so that the behavior of an element p in R is different from its behavior in another relation, R' . If the behaviors in R and R' are not different, there is no

¹⁴ Bertalanffy doesn't explicitly use the term "emergence" in his seminal work on General Systems Theory. However, he does discuss the concept that is now commonly referred to as emergence. He describes it using the phrase "the whole is more than the sum of its parts," which is a fundamental principle in systems thinking and complexity science. In his book (Bertalanffy, 1969, p. 55), he states: "The meaning of the somewhat mystical expression, 'the whole is more than the sum of its parts' is simply that constitutive characteristics are not explainable from the characteristics of isolated parts. The characteristics of the complex, therefore, appear as 'new' or 'emergent'." This description aligns closely with modern definitions of emergence, although Bertalanffy doesn't use the term directly.

interaction, and the elements behave independently with respect to the relations R and R'' .
(Bertalanffy, 1969, pp. 55-56)

Consequently, interactivity is an essential organizational principle that should be considered when studying a system because it gives its unique character and determines the functional roles of its constituent parts.

Interactivity comes with many degrees, and not every degree of interactivity is compatible with life. A system whose components interact to a great extent, where almost everything interacts strongly with each other, is not complex in any meaningful way. Such a configuration would make the system an undifferentiated mix of parts and interactions, where neither structure nor behavior would have probably emerged. Take, for example, the human brain, which consists of nearly 100 billion neurons that are connected in more than 100 trillion synapses. Each time the brain receives a sensory stimulus or produces a movement, various neurons interact to form activation patterns that represent that function. Additionally, when a memory is stored, or new knowledge is acquired, new connections are made, and others are likely to be removed in a process called neuroplasticity. The crucial point is that a brain's function depends not on how much its neurons interact and connect but on how a specific fraction of neurons connect to perform a function. It is not related to quantity but quality. Thus, in the case of a system with many interactive parts, different strengths of interactions (strong, weak, not at all) create structure, viz., subsystems with different functional roles that contribute to the system's overall behavior. Each behavior at a given level comes from interactions of entities at a lower level, and this pattern extends recursively downwards, producing a hierarchy of organizational levels. This hierarchical organization is a common feature in biological complex systems and is often referred to as *nested complexity* (Oltvai & Barabási, 2002). The parts that interact strongly at the lower level will most likely correspond to an operation detected at an upper level. Namely, the upper level's operation is performed jointly by component parts and functions interacting on the lower level.

This leads us to another important organizational feature of living systems: the property of modularity. Modularity in biology highlights the organization and functioning of living systems as composed of distinct yet interconnected units and is a fundamental principle in the study of biological complexity. The tendency of biological systems to consist of distinct, semi-independent units or modules that interact and work together to perform specific functions is widely observed across different levels of biological organization, from the molecular and cellular levels to the level

of whole organisms and ecosystems. Examples familiar to anyone with even a basic understanding of the biological sciences are molecular and cellular modularity: (1) At the molecular level, cells are composed of various functional modules, such as proteins, which can interact with one another to carry out specific tasks. Even more so, proteins often have distinct domains that serve different functions, allowing them to participate in multiple processes within the cell. (2) Within a cell, different organelles act as modules, each with its unique set of functions. For instance, the nucleus houses the genetic material, the mitochondria generate energy, and the endoplasmic reticulum is involved in protein synthesis and transport. These organelles work together to maintain cell function. (3) At an even higher level of organization, modularity is often reflected in the structure of biological networks, such as protein-protein interaction networks or gene regulatory networks. These networks tend to have densely connected subgroups that correspond to functional modules.

In Cognitive Science, one of the most representative expressions of modularity is in the case of language. It is believed that linguistic capabilities are built on a modular structured cognitive system with components that have autonomous roles, the combination of which is what we call language. One of the most common models proposes that language consists of five components (Chierchia, 1999): Morphology, phonology, syntax, phonetics, and semantics. This decomposition can be extended to lower levels, meaning that each of these five components consists of lower-level components that interact and comprise a functional role essential for language production. For instance, words play a central role in the morphology module. A word can be decomposed into morphemes, the smallest components in which a separate meaning can be identified. Morphemes interact based on specific rules, resulting in the formation of several words that enrich language. A similar decomposition can be made to the phonology module, which concerns speech sounds (phonemes) that interact to produce spoken words (Chierchia, 1999). Morphemes and phonemes are components of morphology and phonology, respectively, but we cannot assume that morphemes interact with phonemes because they have different functional roles. Their interactivity is selective and determines the functional organization of the language system.

Modularity is the result of different degrees of interaction between the components of a system. When several parts interact intensively within a group consisting of specific parts while not interacting with other parts of the system, they create a dynamic functional structure within the system. According to the modularity principle, this structure creates functionally specialized and autonomous subsystems called modules. The property of modularity in a system relates closely to the system's near decomposability (Callebaut, 2005), that is, its property to be able to (almost) be

decomposed into subsystems based on the degree of interaction between its components parts. Callebaut stresses that “modules within a system or process are tightly integrated but relatively independent or 'dissociable' from other modules” and adds that “because the strength or weakness of interactions is a matter of degree, modularity should itself be seen as a gradual property” (Callebaut, 2005, p. 6).

Brain networks are a typical example of modularity combined with near-decomposability. As reviewed by Sporns and Betzel (2016), the brain's network architecture is characterized by modules or communities of densely interconnected nodes, with sparser connections between these modules. This modular organization is evident across multiple scales of brain structure and function, from macroscale regions forming distinct functional modules (e.g., sensory processing, motor control, higher-order cognition) to smaller nested submodules. These modules exhibit stronger internal connectivity and coherence in their activity patterns compared to their inter-module connections, allowing for specialized processing within modules while maintaining the ability to integrate information across the entire brain network. Simultaneously, hub regions and weak inter-modular connections facilitate information integration when necessary, striking a balance between segregation and integration. This balance enables the brain to rapidly reconfigure its functional architecture in response to changing task demands or environmental conditions.

Bechtel and Richardson (2010), referring to the explanation of complex biological systems, argue that a system's decomposability hinges on its modular organization, as this is the sole means by which specific functions can be attributed to distinct component parts: In modular organization, “... the components of the system can be subjected to separate study and investigation; it requires that the components have discrete intrinsic functions intelligible in isolation, even if such functions do not independently replicate those of the system as a whole” (p. 25). The crucial point here is whether the whole system can influence how component parts of specific subsystems behave. As long as the interactions remain strong within the boundaries of a subsystem, while the interactions with parts of other subsystems are minimal and there is also little determinative influence from other environmental factors, the subsystem or module remains autonomous. In other words, the “enemy” of modularity is not interactivity per se but the interactivity that crosses the boundaries of a subsystem and the one that is associated with cascading effects, feedback loops, and other effects that obscure the individual role of each element of the system within the subsystem where it is assumed that it functions.

Studying the interactions and the modularity of a system can be utilized for understanding the behavior and implementation of a complex system. The formation of clusters of parts due to intense interactions between them indicates the existence of subsystems or modules that perform a particular function. A relevant biological example can be found inside the muscle fibers' sarcomeres, where there is an abundance of filaments formed by two proteins, actin, and myosin. In this example, the forming of functional subsystems is defined dynamically through various degrees of interactivity between the system's components. Myosin filaments develop projections that are called *cross-bridges*, and they interact with actin filaments. The intensity of interactions between these two entities varies in different stages of the contraction mechanism. In the non-contraction condition, there are weak interactions between myosin and actin: myosin heads continuously sample weak bindings with several actin-binding sites. But when myosin molecules consume ATP energy and their projections enter the appropriate angle to exert traction, the interaction's profile changes. Strong binding sites are exposed in actin, they immediately bind strongly with myosin heads, and this triggers the *power stroke* phase of the mechanism, namely the pulling of myosin towards the center of the sarcomere, which produces the contraction of the muscle fiber (Guyton & Hall, 2006; Davis, Tikunova, & Janssen, 2019; Oatis, 2009). Weak interactions characterize the system as being in a relaxed standby state of readiness, while the strong ones organize and transform it into a machine. Strong interactions suggest that specific components work together to perform a function; they are a functional subsystem within the muscle cell. Detecting the strength of the interactions between the parts of a system and grouping them so that functional subsystems are formed is a heuristic method to unravel its functions.

2.4. IDEALIZATION AND NEAR-DECOMPOSABILITY: BRIDGING SIMPLICITY AND COMPLEXITY

The behaviors and functions of biological organisms are multifaceted, interdependent, often difficult to identify, and can be observed at various levels of organization. A fundamental principle for their study is that when many entities interact extensively, they likely work together for a specific function. However, this doesn't preclude weaker interactions these entities may have with other elements within the system. These weaker interactions may contribute to other functional roles, making the decomposition of the system into functional subsystems a complex endeavor. This complexity is common in the study of many scientific phenomena.

Our world is incredibly rich with interactions, to the extent that one might assume everything is interconnected and if we attempt to focus equally on all these interactions, we will fail to comprehend it. However, some elements are more closely related than others, and often, to understand a function or phenomenon, it's necessary to ignore less intense interactions. Scientists frequently simplify mechanisms and processes deliberately to construct more tractable models that are easier to understand. This strategy, known as *idealization*, is a common approach in developing explanatory and predictive models of phenomena¹⁵ (Roman & Hartmann, 2018). Idealization involves simplifying or abstracting a complex system or phenomenon to create a more manageable and understandable representation. This process requires making assumptions and disregarding certain details considered less relevant for the specific analysis or model being constructed. Idealization allows scientists to focus on the essential aspects of a system, gain insights, and make predictions more effectively.

As part of the effort towards idealization, the study of interactions and the formation of subsystems within a biological system can be accomplished by treating the system as if it were, ideally, completely decomposable, thereby disregarding weaker interactions. The idealization strategy we will explore here is called *aggregation of variables*, first proposed by Simon and Ando (1961) and later by Ando and Fisher (1963) in the field of economic systems. The idea extended to the field of philosophy of biology (Wimsatt, 1972; 1986), cognitive science (Simon, 1995), and computer science (Courtois, 1977). Courtois has further implemented this technique to study the dynamic behavior (or performance) of large computing systems. His motivation related to the simulation models often used to analyze large and highly complex computer systems. These models, especially when they simulate systems with a high degree of complexity, require a correspondingly large degree of detail that nearly approaches the complexity of the system under study. This evokes many practical problems. Alternatively, someone could use a somewhat simplified approximate model based on a decomposition into constituents that could be separately understood and analyzed and, at the same time, estimate the required degree of approximation. The general idea of the aggregation of variables is the following:

¹⁵ There are two general kinds of idealizations: Aristotelian and Galilean. In the Aristotelian idealizations, the objective is to focus on a limited set of properties in isolation and this is achieved by ignoring all the factors that we do not consider to be directly related to the phenomenon under consideration. In the Galilean idealizations, scientists build models that involve deliberate distortions of the phenomena in order to simplify them without significantly affecting the very essence of the phenomenon. (Roman & Hartmann, 2018)

In many large systems all variables can somehow be clustered into a small number of groups so that: 1) the interactions among the variables of each single group may be studied as if interactions among groups did not exist, and 2) interactions among groups may be studied without reference to the interactions within groups" (Courtois, 1977, p. 4).

This suggestion encompasses two types of idealizations: first, considering the system decomposable by ignoring all the weak interactions, and second, studying each level separately to understand the interactions between its entities isolated from the lower-level entities. To understand this idea in a more illustrative way, let us consider a complex system that we want to study its behavior. Based on the former suggestions about the aggregation of variables, we assume a hierarchical structure consisting of a system composed of subsystems, composed of sub-subsystems, and so on where: (1) the interactions between the variables within each subsystem can be studied as if there are no interactions with variables in the other subsystems and (2) the interactions between subsystems can be studied without any reference to the interactions that exist within subsystems. Such an idealized hierarchical system is completely decomposable and consists of independent subsystems that can be separately studied, as is depicted in *Figure 3*. In this system, the strong interactions of the parts are studied in isolation within the subsystem they constitute. Accordingly, the interactions between the subsystems at the above hierarchical level are studied in isolation without reference to any strong or weak interactions existing between the parts that reside on the lower hierarchical level.

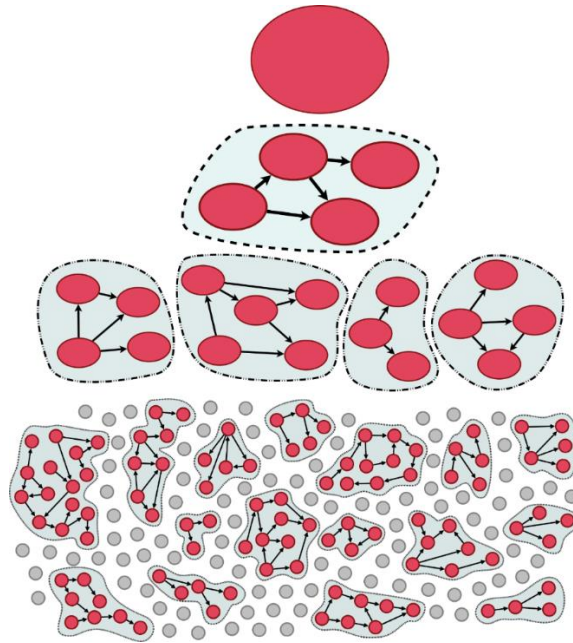


Figure 3: The idealized case of a completely decomposable system. The interactions between entities at each level are confined within subsystems, and no interactions occur with free entities or entities that belong to another subsystem. The circles at the lowest level represent entities. The strong interacting entities (pink circles) form functionally specialized groups that constitute functional subsystems (modules). Each subsystem (enclosed gray areas in the bottom level) corresponds to a pink ellipse at a higher level and interacts with other pink ellipses. Based on these interactions, they form subsystems at this level, too (represented by a dashed line enclosing a gray area). This process repeats recursively up to the top level, where we can see the system that all mentioned entities constitute. However, in most complex systems, it's very difficult to find entities that interact all and only with specific entities, so this situation represents an idealization. It's based on the assumption that due to differences in the intensity or frequency of interactions, it would be a safe choice to only consider the strongest ones and, based on these, to decompose the system, identifying a specific modular structure.

In a complex system with the former characteristics, we can draw inferences for its behaviors when observing that some variables move together in an aggregate way. This means that we decompose it into smaller components. And how can we analyze the overall behavior through this decomposition? A system is an ensemble of elements that produces specific behavior as a whole. Behavior could be something so simple as deciding to go in one direction or another. In this case, we could ascribe to this behavior a variable with two values, 1 for right and 2 for left. The system is composed of parts that could be systems as well; let us call them subsystems, and each of these produces a behavior (as a whole), too, and we can also ascribe variables to them. Similarly, the subsystems are probably composed of smaller parts that produce other behaviors that correspond to variables and so on. The system and the smaller subsystems that constitute it are consequently aggregations of variables, and this is the same as we go down to the elementary parts. They are defined as sub-systems due to their tight connectivity, and simultaneously, they are functionally defined as sub-systems because each one is responsible for a particular behavior. The crucial point is that the interactions between the components in every subsystem are very strong, while there

are no interactions between components of different subsystems, and this interactivity profile reflects the functional differentiation between the sub-systems. This results in a modular organization. A phenomenon observed on the top level of such a system can be explained and predicted by decomposing and localizing it into functions and components.

How far is the nature of biological systems from the ideal state of absolute modularity, where a system is completely decomposable? Complex systems, as we have seen, often exhibit emergent properties, which are characteristics or behaviors that arise from the interactions of individual components but cannot be easily predicted or deduced by examining these components in isolation. These emergent properties are a hallmark of complex systems and contribute to their overall complexity. When attempting to decompose a complex system into its components or modules, there is a risk of oversimplifying or overlooking these emergent properties and interactions. Decomposing a complex system to the point where it can be fully understood solely by examining its parts may result in losing the holistic understanding of how parts interact to give rise to the system's behavior, given that in these systems, modular and global properties are combined.

Modularity and complexity often coexist and interact in biological systems. Paradoxically, modularity can actually enhance the complexity of a biological system by allowing it to perform diverse functions through the coordination of these modules. Each module may have its own complexity, contributing to the overall complexity of the system. Consequently, biological organisms are complex, hierarchical, and evolve over time, making the notion of complete decomposition inapplicable to them. Only linear systems are fully decomposable. In contrast, nonlinear systems are non-decomposable and can only be characterized using global collective variables that summarize the behavior of all the system components (Chemero, 2009). In these systems, interactions operate across multiple scales of time and size, influencing the system in many ways.

A system that combines the above characteristics is near-decomposable (Simon & Ando, 1961; Simon, 1962; Simon, 1995). Near-decomposability is a way to describe hierarchically organized systems that are not fully modular but have a degree of modularity that allows for more manageable analysis without sacrificing the interactions necessary for emergent properties. Interestingly, it is possible to study them in both ways, as if they have a modular and global nature simultaneously. Herbert Simon defines them as systems arranged in a hierarchy of components and subcomponents, with interactions within components that are strong and fast compared with the

interactions between components at the same level (Simon, 1962). A near-decomposable system can be broken down into subsystems based on the strength of causal interactions between its constituent parts, resulting in clustered subsystems with strong interactions contained within them. The accuracy of these subsystems depends on the contrast between strong and weak interactions. In such a system, we can ignore weak interactions and focus on what happens only with the strong ones.

With Ando (Simon & Ando, 1961), they showed that in these systems the analysis of the short-run and that of the long-run dynamics should be separated. They proved two theorems: The first one states that, provided intergroup dependencies are sufficiently weak as compared to intragroup ones, in the short run, the system will behave approximately and can, therefore, be analyzed as if it were completely decomposable. The second theorem states that even in the long run when neglected intergroup dependencies have had time to influence the system behavior, the values of the variables within any group will remain approximately in the same ratio as if those intergroup influences had never existed. The results obtained in the short run will, therefore, remain approximately valid in the long run, as far as the relative behavior of the variables of the same group is concerned (Courtois, 1977). This means that in the long run, we will have robust conclusions both for the outcomes of interactions within each subsystem and for the overall properties that the system acquired due to the weak interactions that occurred among variables belonging to different groups. Treating the system as decomposable can capture both its modular and global nature.

An illustrative example of complex system behavior is a thermally insulated building (an isolated system¹⁶) with numerous large rooms that act as moderate insulators (major subsystems) and smaller cubicles that function as pure insulators (Simon & Ando, 1961; Simon, 1962). Let us consider a scenario where we measure the temperature at time t_0 , observing a wide variation in temperature between rooms and cubicles - a state of thermal disequilibrium. After several hours, at time t_1 , we would still find significant temperature variations among rooms. However, after several days, we would observe an almost uniform temperature throughout the building. This equilibration process results from the aggregate behavior of all the system's subcomponents over time, viz., in

¹⁶ In thermodynamics, an isolated system does not exchange energy or matter with its surroundings. It is completely self-contained and does not interact with its environment in terms of energy or mass transfer. It's a theoretical construct used for simplification and analysis in thermodynamic studies.

the long run (*figure 4*). Initially, by deliberately neglecting small-scale interactions (heat diffusion) between the large rooms and emphasizing large-scale interactions among smaller cubicles, we obtain an approximately accurate result: each individual room exhibits its own distinct temperature, a property specific to each separate room. Ultimately, the smaller-scale interactions lead to a new uniform temperature, which becomes a property of the entire system. The provided example highlights several key aspects of complex, near-decomposable systems. It demonstrates the importance of scale in system interactions and illustrates the emergence of system-wide properties from local interactions. The example also emphasizes the role of time in the evolution of system behavior and showcases the transition from local properties to global properties. Apparently, the influence of time on these dynamic systems should always be considered upon their examination.

	A1	A2	A3	B1	B2	C1	C2	C3
A1	-	100	-	2	-	-	-	-
A2	100	-	100	1	1	-	-	-
A3	-	100	-	-	2	-	-	-
B1	2	1	-	-	100	2	1	-
B2	-	1	2	100	-	-	1	2
C1	-	-	-	2	-	-	100	-
C2	-	-	-	1	-	100	-	100
C3	-	-	-	-	2	-	100	-

A1	B1	C1
A2		C2
A3	B2	C3

A hypothetical nearly-decomposable system. In terms of the heat-exchange example of the text, A1, A2, and A3 may be interpreted as cubicles in one room, B1 and B2 as cubicles in a second room, and C1, C2, and C3 as cubicles in a third. The matrix entries then are the heat diffusion coefficients between cubicles.

Figure 4: A nearly-decomposable system illustrating heat diffusion in a thermally insulated building. Reproduced from Simon (1962): a hypothetical nearly-decomposable system using the example of heat diffusion in a building with three rooms, each containing multiple cubicles. The matrix represents heat diffusion coefficients between cubicles: (1) Room A: Cubicles A1, A2, A3, (2) Room B: Cubicles B1, B2, (3) Room C: Cubicles C1, C2, C3. The matrix entries indicate the strength of heat diffusion between cubicles: 100: Strong heat diffusion (between cubicles in the same room), 1 or 2: Weak heat diffusion (between cubicles in adjacent rooms), Dash (-): Negligible or no direct heat diffusion. This system demonstrates the principle of near-decomposability, where interactions within subsystems (rooms) are stronger than interactions between subsystems. Despite

the varying diffusion rates, the entire building will reach thermal equilibrium over time, with all rooms and cubicles attaining the same temperature after approximately one week.

More specifically, the general way of thinking in analyzing nearly decomposable systems is as follows (Simon & Ando, 1961; Simon, 1962; Courtois, 1977): A nearly decomposable system is a complex system that can be represented by a hierarchical structure of interacting components. To understand and analyze such a system, we can use matrix representation. Let's consider a system with three levels of interaction: P, Q, and R.

$$P = \begin{bmatrix} P1 & Q1 & R1 & R2 \\ Q2 & P2 & R3 & R4 \\ R5 & R6 & P3 & Q3 \\ R7 & R8 & Q4 & P4 \end{bmatrix}$$

In this matrix:

- P₁, P₂, P₃, P₄ represent strong interactions within subsystems (diagonal elements)
- Q₁, Q₂, Q₃, Q₄ represent intermediate-strength interactions between subsystems
- R₁-R₈ represent weak interactions between subsystems

We can analyze this nearly decomposable system on three time-scales: (1) In the short-term analysis, we focus on the strong interactions (P elements) and treat the system as if it were completely decomposable by ignoring Q and R elements. This gives a first approximation of system behavior. (2) In the intermediate-term analysis, we include the Q elements to account for moderately strong interactions between subsystems in order to refine the short-term analysis. (3) Lastly, in the long-term analysis, we consider all elements (P, Q, and R) to understand the system's behavior over extended periods. The weak R interactions, while individually small, can have significant cumulative effects over time. If it is desired to understand interactions at different scales, we could conduct a multi-level analysis by studying each level (P, Q, R) separately, and to simplify the analysis, we could group variables that move together (aggregation of variables). By applying these steps, we can gain insights into both the modular nature of the system (through P elements) and its emergent properties (through Q and R interactions), that is, an understanding of the system's behavior at different time scales and levels of organization. *Figure 5* illustrates and explains the differences between a completely decomposable and a nearly decomposable system by examining the matrices that represent the interactions of the subsystems' components.

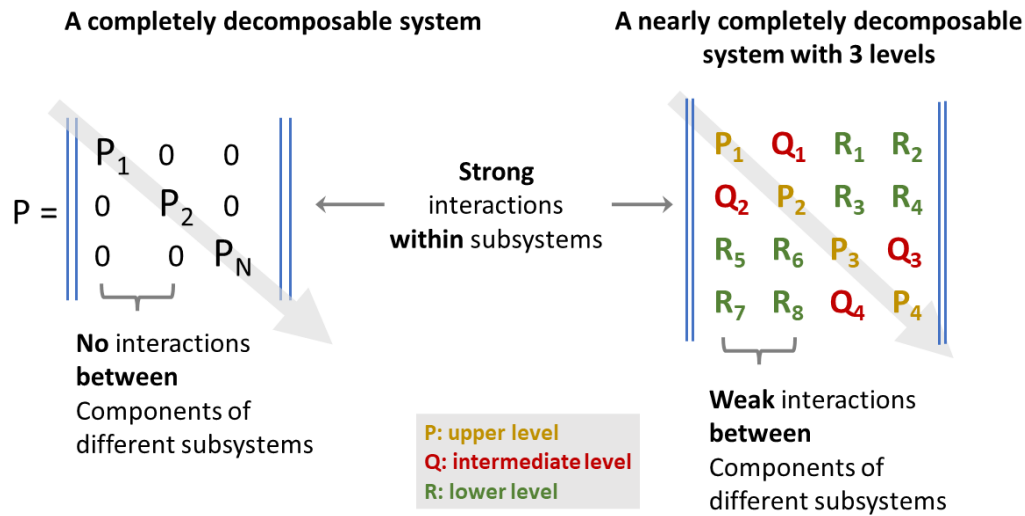


Figure 5: Difference between a completely decomposable system and a near-decomposable system. Based on data from Simon and Ando (1961), Simon (1962) and Courtois (1977).

Left side: A completely decomposable system. In this system, the matrix P shows strong interactions (represented by $P_1, P_2, \dots, P_N$) along the diagonal, which correspond to the interactions within subsystems. All other elements in the matrix are zero, indicating no interactions between components of different subsystems.

Right side: A near-decomposable system with three levels of interaction. The upper level (P) still shows strong interactions along the diagonal, similar to the completely decomposable system. However, this system also includes two additional levels of interaction: (1) An intermediate level (Q), represented by elements Q_1, Q_2, Q_3 , and Q_4 . (2) A lower level (R), represented by elements R_1 through R_8 . These additional levels (Q and R) represent weaker interactions that exist between subsystems. While these interactions are smaller in magnitude compared to the strong interactions within subsystems (P), they are not negligible. Their presence is what makes the system "near-decomposable" rather than "completely decomposable." In a near-decomposable system, the interactions above and below the diagonal are small but significant. They allow subsystems to influence each other and contribute to the overall behavior of the system in an aggregate way. This crucial characteristic means that while we can still meaningfully decompose the system into subsystems for analysis, we cannot completely ignore the interactions between these subsystems.

The concept of near-decomposability is of immense importance for the ability to study and analyze many biological systems that are thought to exhibit this property. Beyond the examples already mentioned, one can recognize the principles of near-decomposability in most biological structures that simultaneously display functional specialization and modularity. This is due to high connectivity between specific biological units and system properties that arise from weaker interactions between entities belonging to different modules. For example, *metabolic networks* exemplify near-decomposability through their hierarchical organization and modular structure. Research by Ravasz et al. (2002) has shown that these networks exhibit scale-free topology and high clustering coefficients, indicating significant local clustering. This structure is characterized by

small, highly connected clusters organized hierarchically into larger, less cohesive units, a pattern consistent across organisms from bacteria to complex eukaryotes. The modular architecture allows for functional independence among different modules while maintaining overall network interconnectedness, facilitating efficient regulation, adaptation, and response to environmental changes. Importantly, these modules often correspond to known biological functions or pathways, suggesting that this organization has functional significance in metabolic system operation and evolution. This hierarchical and modular nature of metabolic networks provides a clear example of near-decomposability in biological systems, enabling efficient functioning, adaptability, and evolvability across diverse organisms.

Another interesting example that exhibits near-decomposability regards the so-called *transcription regulatory networks (TRNs)* which serve as sophisticated biological systems controlling gene expression in organisms (Alon, 2007). These networks, crucial for cellular responses to environmental stimuli and regulation of various biological processes, are composed of complex interactions between genes and regulatory proteins, primarily transcription factors. TRNs are characterized by their modular structure, featuring recurring patterns of interactions called *network motifs*, which act as fundamental building blocks and perform specific functions. The motifs, observed across diverse organisms from bacteria to humans and often evolutionarily conserved, contribute to the network's hierarchical structure, where some transcription factors regulate others, creating cascades of gene regulation. The near-decomposable nature of TRNs is evident in these network motifs since they function both as modules with stronger internal interactions and with weaker connections to other parts of the network. This modularity allows researchers to analyze individual motifs' dynamics even within the broader system, while the hierarchical organization creates layers of regulation that can be studied at different abstraction levels.

In summary, this final section aimed to highlight the fact that even in a system far removed from an ideal state that would render it completely decomposable and thus analyzable using decomposition-based strategies focusing on components or functions, we can still apply functional or componential decomposition to obtain results that explain its behavior to a satisfactory degree of approximation. This conclusion holds for systems generally considered near-decomposable. Such systems typically exhibit the following characteristics: Hierarchical structure, as described by Simon (1962, p. 468), namely “composed of interrelated subsystems, each of the latter being, in turn, hierarchic in structure until we reach some lowest level of elementary subsystem.” This

physical characteristic is evident through the observation of distinct structural or functional units at different scales. Secondly, modularity; as we have seen, near-decomposable systems display identifiable modules or subsystems with stronger internal interactions compared to interactions between modules. This property can be observed in clustered components or functions with clear boundaries. A third typical feature is time scale separation, as demonstrated in the building example, which refers to the existence of different rates of change or response times for processes within subsystems compared to interactions between subsystems. This separation contributes to the robustness of subsystems, allowing them to maintain function when other subsystems are perturbed, a feature examined by Kitano (2004) in the context of biological robustness¹⁷. Nearly decomposable systems also exhibit emergent properties (Bar-Yam, 2019), viz., system-level behaviors that arise from the interaction of subsystems but are not predictable from subsystems alone. Physically, this is often seen in complex feedback loops or non-linear interactions between subsystems. The human immune system exemplifies this, where the coordinated action of various cell types and molecular signals leads to emergent behaviors like adaptive immunity.

There are various other features observed and studied in nearly-decomposable systems that are either based on or strongly related to and dependent on those mentioned in this section. Near-decomposability in information flow patterns, which is seen in specialized communication channels or signaling pathways within modules, is one such case. In neural networks, for instance, we

¹⁷ Biological robustness is a fundamental characteristic of living systems that enables them to maintain functionality despite internal and external perturbations. This property is achieved through various mechanisms, including system control, alternative pathways, modularity, and decoupling. Robustness is closely linked to evolution and evolvability, as it allows organisms to adapt to changing environments over time. The architecture of robust biological systems often involves genetic buffering, robust modules, and bow-tie structures, which contribute to their ability to withstand environmental changes while remaining genetically flexible (Kitano, 2004).

In brain, robustness is a very common characteristic and is expressed mainly through neural plasticity, which involves the brain's ability to reorganize itself by forming new neural connections throughout life enabling adaptation to new experiences, recovery from injuries, and learning. It is also expressed through redundancy in neural networks, which refers to the existence of multiple pathways and circuits in the brain that can perform similar functions, providing robustness against localized damage (Tononi, Sporns, & Edelman, 1999). Other expressions of robustness in brain include hemispheric specialization and cooperation (Gazzaniga, 2000), the brain's resilience to neuropathological damage and homeostatic plasticity, that is, the brain's ability to maintain stable neural activity despite changes in input or internal state (Stern, 2009; Turrigiano & Nelson, 2004). Robustness can be observed in cognitive functions too. The brain's ability to process information across multiple regions, providing robustness against localized disruptions (distributed information processing) (Bullmore & Sporns, 2012) and to adapt cognitive processing strategies to face new and unexpected conditions in the environment (cognitive flexibility) (Dajani & Uddin, 2015) are some of those.

observe dense local connectivity within brain regions, with sparser long-range connections between regions (Tononi, Sporns, & Edelman, 1994). Evidently, the physical characteristics that make a system nearly decomposable are multifaceted and interconnected, allowing for both the relative independence of subsystems and their integration into a coherent whole. This enables complex systems to achieve both efficiency and adaptability. Understanding these physical manifestations of near-decomposability is crucial for analyzing and modeling complex systems across various scientific disciplines.

2.5. SUMMARY AND CONCLUSIONS

In this chapter, I aimed to provide insights into the organization of biological structures, from individual organs to entire organisms, offering a framework for unraveling the multi-level complexity inherent in living systems. In the realm of complex biological systems decomposition is a key tool to unpack their multi-level complexity. My exploration ranged from the basic ideas of composition and decomposition to the interplay between aggregativity and emergence and the fine balance of interactivity and modularity that shapes biological structures.

I began by recognizing that biological systems are typical examples of complexity, with many interacting parts across various levels of organization. Decomposition is a versatile strategy that can be applied to both physical structures and abstract functions. This flexibility allows researchers to reveal various layers of biological complexity, showing how parts relate to wholes and how structure relates to function. A key idea is the important role of restricted forms of aggregativity in shaping the hierarchical organization of biological systems and the role that sub-aggregativity plays in forming nested hierarchies, allowing higher-level structures and functions to emerge from lower-level components. This principle helps explain how biological systems stay organized across different scales, from the interactions of molecules to the dynamics of ecosystems.

The discussion of interactivity and modularity highlighted another crucial aspect of biological organization: the strength and nature of interactions between components can lead to the formation of functional modules - semi-independent units within the larger system. This modular organization is a key feature of biological complexity, allowing for specialized functions while maintaining the flexibility needed for adaptation and evolution. Hence, there is a delicate balance between decomposability and emergent properties in biological systems. Notably, while

decomposition helps us simplify and understand complex systems, we should be careful not to oversimplify and lose sight of the emergent behaviors that are essential to how the system works. The concept of near-decomposability, initially proposed by Herbert Simon (1962), offers a valuable middle ground, allowing for idealized analysis while still capturing both modular and emergent whole-system properties.

I conclude the final section of the chapter with an extensive exploration of the concept of near-decomposability, starting from its initial conception and ending with its application to many biological cases. This particular section is very important for forming the present thesis because it sets out the reasoning by which, even if a system cannot be fully decomposed through a functional or structural decomposition, it can nevertheless be studied with explanatory tools that presuppose decomposability, as long as it shows characteristics of a nearly-decomposable system. This will enable us to explore, in the second part, specific aspects of how mechanistic explanations apply to cognitive phenomena, employing a more flexible decomposability condition for systems that exhibit the mental behaviors under study.

Lastly, concluding this chapter it should be appreciated that the study of complex biological systems through decomposition highlights the importance of looking at both the details and the big picture to recognize the complexity that arises from seemingly simple interactions.

CHAPTER THREE - EXPLORING BIOLOGICAL MECHANISMS VIA COMPOSITIONAL EXPLANATIONS

3.0. OUTLINE

Having already investigated the two basic concepts that form the axis of this dissertation, viz. levels, and decomposition, I explore here the idea of compositional explanations in biological systems and demonstrate its application across a hierarchy of levels with a case study in human kinesiology. I begin presenting how explanatory models in biological sciences evolved (3.1. *The Evolution of Explanatory Models in Biological Sciences*) and examine the limitations of traditional nomological (3.1.1.) and statistical (3.1.2.) approaches to explanation in biology, demonstrating why these methods often fall short when applied to complex biological systems. I then introduce the concepts of decomposition and multilevel analysis as foundational elements for understanding biological mechanisms (3.1.3.), setting the stage for a more comprehensive approach to explanation in the life sciences. The role of decomposition is further demonstrated through Gillett's (2016) work on scientific composition (3.2. *Gillett's Framework on Scientific Composition*), which offers a detailed account of how compositional explanations can be constructed and applied in biological contexts. I then, apply Gillett's framework in a detailed case study of an elbow movement (3.3. *Elbow Movement as a Model for Multilevel Biological Analysis*) to demonstrate how a complex biological system can be analyzed and explained across multiple levels of organization. The case study begins with an overview of the anatomical and physiological foundations of forearm flexion (3.3.1.) and subsequently, I systematically decompose the elbow movement system to reveal its hierarchical organization, illustrating how different levels of explanation emerge through this process (3.3.2.). Finally, I focus on muscle contraction mechanisms and examine them through the lens of Gillett's scientific composition (3.3.3.), demonstrating how this approach can provide a rich, multi-level understanding of biological phenomena.

3.1. THE EVOLUTION OF EXPLANATORY MODELS IN BIOLOGICAL SCIENCES

Explanation is both the primary pursuit and product of science. Salmon (1999) describes scientific explanation as an endeavor of using some particular or general facts drawn from empirical

sciences, aiming to make a particular event or general fact intelligible and understandable. An explanation may answer a question about why or how something happens and is contrasted to confirmation or verification, where science proves that something actually happens. Undoubtedly, what constitutes a scientific explanation is a major philosophical question that had preoccupied scientists and philosophers since antiquity. As of today, the modern debate and related disagreements were revived in the middle of the last century when Hempel and Oppenheim proposed their *Deductive-Nomological (DN)* Model of explanation (Hempel & Oppenheim, 1948). The *DN* model became a milestone in the later treatment of the issue since all other explanation's accounts that emerged afterward, either as new considerations or as developments of older ones, were positioned around it. In any case, all proposed models of scientific explanation begin from the following common starting points, which set the basis for any conversation.

Firstly, in an explanation, there is always an *explanandum*, that is, a particular event or a general fact that we want to explain, and an *explanans*, which consists of the propositions, the relationships, the facts, or more generally, the theory that explains and predicts it, that is, it describes why it occurs. Secondly, a scientific explanatory model should account for the nomological status of the explanans. This is essential for being generalizable and nomologically valid. Thirdly, a scientific explanatory role is served when a theory or model explains a phenomenon, not by simply describing its possible evolution over time, but by examining the causal relationships involved. These three starting points have been implemented in different ways in the various accounts of explanation that emerged in the recent philosophical discussion.

After the initial model of Hempel and Oppenheim, which was widely criticized, Carl Hempel (1962; 1965) extended his original account by adding the element of statistics in the explanation, resulting in the *Inductive-Statistical (IS)* and *Deductive-Statistical (DS)* models of explanation. The importance of statistical correlation was further emphasized by Wesley Salmon (1971) in his *Statistical Relevance (SR)* model. However, he later revised this first model and suggested his *Causal Mechanical (CM)* model of explanation (Salmon, 1984; 1994) that interpreted causal relationships as processes. Cummins, meanwhile, had already begun to develop his explanatory scheme of *Functional Analysis* (Cummins, 1975; 1983). Instead of focusing on laws that govern the phenomena under explanation, his approach, tailored for the psychological domain, was based on systemic analysis and decomposition methods.

The emphasis on the decomposition of complex hierarchical systems was already proposed by Herbert Simon (1962) and later developed by Kaufmann (1970) and Wimsatt (1972), both focusing mainly on biological phenomena. A blending of the former ideas, which stressed the importance of causal processes and decomposition, with a revival and reappraisal of the old mechanistic philosophy, led to the development of a new account adapted to the particular nature of the biological phenomena. It was named *New Mechanical Philosophy* (Machamer, Darden, & Craver, 2000) and began to dominate in explaining phenomena related to biological complexity. In this account, central is the role of a system's recursive decomposition and the search for an explanation in a hierarchy of levels. The core idea is the notion of *compositional explanation*, which will be presented in this chapter after a more detailed review of the philosophical debates that led to its development.

3.1.1. From Laws to Mechanisms: Shifting Paradigms in Biology

Initially, laws were proposed as being key to explanation. The recent argumentation on the subject starts with a model that conceptualizes the explanatory process as a deductive argument proving the occurrence of the explanans by combining a law of nature with specific initial conditions. This summarizes the influential Deductive-Nomological (*DN*) model of Hempel and Oppenheim (1948) that laid the foundations for a discussion which, although it ended up challenging its applicability in many scientific domains, nevertheless highlighted the essential role that laws play in scientific explanation. The general form of an explanation that follows the *DN* model consists of several sentences playing the role of the explanans, which necessarily lead to a conclusion that plays the role of the explanandum. A good explanation has a sound argument's structure, with valid premises, the latter provided that their truth draws from empirical facts. As Hempel and Oppenheim suggested, there are three logical and one empirical condition of adequacy for an explanation:

I. Logical conditions of adequacy.

(R1) The explanandum must be a logical consequence of the explanans; in other words, the explanandum must be logically deducible from the information contained in the explanans, (...)

(R2) The explanans must contain general laws, and these must actually be required for the derivation of the explanandum.(...)

(R3) The explanans must have empirical content; i.e., it must be capable, at least in principle, of test by experiment or observation.(...)

II. Empirical condition of adequacy.

(R4) The sentences constituting the explanans must be true. (Hempel & Oppenheim, 1948, p. 137)

An essential aspect of this model, particularly emphasized by Hempel and Oppenheim, is that such a formal analysis is suitable not only for a scientific explanation but also for predictive purposes.

What makes the difference, then? If the antecedent conditions and laws are known, then the phenomenon is derived from them as a prediction. In the case of explanation, though, the explanandum phenomenon is a given, and the scientific inquiry focuses on the discovery of the sentences, being laws or initial conditions, which explain it. There is an apparent symmetry between prediction and explanation. To summarize the whole idea, the argument schema that Hempel and Oppenheim proposed is a logical deduction from the explanans, i.e., the statements of antecedent conditions, and the general laws to the explanandum, i.e., a description of the empirical phenomenon to be explained.

In this account, the role of laws involved in the argument that serves the explanatory purpose is so essential that the model is often referred to as the *Covering Law (CL)* model. One or more premises in the argument are *lawlike sentences*, which are of two types: general statements that are analytic and lawlike scientific sentences with empirical content. A lawlike sentence in a deductive explanation, apart from being true, must satisfy specific conditions: it should have a universal form, such as “all *X* are *Y*,” or a conditional universal form, such as “at conditions *C*, all *X* are *Y*.”

Additionally, it should refer to a fundamental law with non-limited scope, or it can be derivative. Derivative means that apart from being of universal character, it follows from some fundamental laws. A fundamental law is a law of nature, and as such, it has two essential characteristics (Salmon, 1999). First, it supports counterfactual inferences. Counterfactual inferences are formulated as hypothetical propositions of the type “if *X* happened, then *Y* would happen.” The law assures their validity. Second, it supports modal statements of physical necessity and impossibility; that is, it describes what is possible in the world based on the law and what is not.

A collateral version of the model was developed at the same time by Nagel, focusing on biological phenomena. Nagel (1961; 1977) proposed a theory reduction account whose basic idea is that one theory reduces to another when the reduced theory is derivable from the reducing theory, and there are some bridge laws between them. He considers his account as a *DN* explanation where the

explanandum is the reduced theory, and the explanans is the reducing theory. The explanans logically implies the explanandum, so here, too, the explanation has an argument form. Nagel's support for the *DN* model also extends to explaining biological functions, where a function such as the cardiac can be formulated as an explanatory argument (Nagel, 1961). According to Nagel, someone can transform the description of a function into a deductive argument, and it can be done in the following way:

"The function of *A* in a system *S* with organization *C* is to enable *S* in environment *E* to engage in process *P*. And this can be expanded into an explicit explanation in this way: Every system *S* with organization *C* and in environment *E* engages in process *P*. If *S* with organization *C* and in environment *E* does not have *A*, then *S* does not engage in *P*. Hence, *S* with organization *C* must have *A*." (Nagel, 1961, p. 403).

Let's consider the following question: *why* does the heart pump blood? The function of the heart is to pump blood to the vessels to supply the body with oxygen. This description of the heart's function can be enriched with other internal and external clues, such as that every vertebrate, being a physical organism, along with other internal elements, possesses a heart and lives in an external environment that is rich in oxygen. For Nagel, this can be reformulated as a deductive argument of the following form: Every vertebrate organism with a particular internal structure, living in a specific environment, needs oxygen to survive. To ensure this, the heart must send oxygen to the tissues. If an organism with this internal structure living in this environment lacks a heart, it cannot deliver oxygen to the tissues. Therefore, the organism must have a heart. However, this account doesn't explain what the heart does; rather, it proves the necessity of its existence. It assumes that functions are linked to specific goals of the larger system. This represents a different interpretation of the explanatory question "*why*": it doesn't address *how* a function is performed, but instead justifies the purpose of its existence. Another important issue highlighted regarding this type of explanation for biological phenomena relates to their transitive nature, i.e., there is a transition from one state to another. As Cummins (1983) emphasizes, an adequate explanation should refer to the causal role of the organ's function as part of a more general bodily capacity. Indeed, this is not the case in nomological explanations, which focus primarily on the logical form of sentences describing the occurrence of a phenomenon.

Furthermore, significant problems arise when attempting to explain phenomena involving complex system behaviors, whether using the *DN* model or the related Nagelian reduction model. These

phenomena often require analysis at multiple hierarchical levels, as is the case with biological systems. To study the role of cardiac function in the human body, one must examine it at the organ level, considering the heart as a muscular pump that sends oxygenated blood to the periphery (systemic circulation) and receives deoxygenated blood to send to the lungs for reoxygenation (pulmonary circulation). However, this is only one part of the story, described at a relatively high level of organization. A complete explanation must also address oxygen's role in each circulatory path, which requires analysis at a much lower cellular level. For a more comprehensive understanding of the heart's function in the organism, the oxygen enrichment mechanism in the lungs' alveoli is essential. Consequently, such an explanation spans various hierarchical levels of organization and, to be complete, necessitates decomposing the phenomenon and the system that instantiates it into numerous component parts and functions described at various compositional levels. It seems unlikely that explanatory deductive arguments alone can adequately capture this kind of hierarchical complexity.

Indeed, Craver (2007) emphasizes that explanations of this kind are detached from the methodologies of biological sciences, while Gillett (2016) expresses an even more rigorous and broader critique regarding the methodological alienation of both the *DN* model and the relative Nagelian reduction. Considering the common scientific practice that investigates phenomena at many levels of decomposition, Gillett argues that these explanatory approaches are far from what scientists do and consider high-level sciences dispensable, making lower sciences, at least in principle, explanatory omnipotent. Furthermore, by not accepting compositional explanations, which scientists consider valid and successful, they are outdated and not influenced by modern scientific discussions.

Suppose now that we accept an explanation of only the lower-level behaviors by looking at the laws and antecedent conditions that conclude to them. Even in that case, it would be difficult to identify the kind of laws that Hempel and Oppenheim require in constructing the deductive argument they propose. The intense critique of Craver (2007) on the applicability of deductive explanation for neurobiological phenomena reveals that such explanatory schemata are more appropriate for the physical sciences rather than for biology since, in the biological field, it is difficult to distinguish between nature's laws and accidents or non-explanatory generalizations. Besides, the role of antecedent conditions in biological phenomena is much more complicated, as the distinction between what constitutes a genuine cause for a phenomenon and what is classified as enabling and background conditions is often quite vague. Gillett (2016) adds that it is often the case that

explanatory relevant phenomena do not make the explanandum more probable; on the contrary, sometimes explanatory factors reduce the probability for it to occur. Hence, an explanation like the one described in the *DN* model, apart from insufficient, seems quite strict and impossible for these phenomena.

Bechtel (2008) expresses a similar critique. He stresses that nomological explanations draw from the fact that nature follows regular patterns and that a sufficient explanation shows that an event falls under such a pattern. However, laws of nature have a broad scope and appliance. On the contrary, generalizations in the biological sciences are more specific. Furthermore, the explanations that someone finds in the biological field focus on the composition and organization of behavior-producing mechanisms. So, nomological explanations are far from the methodologies and theories of the biological sciences. They are a legacy of analytic philosophy that analyzes concepts in terms of their necessary and sufficient conditions. Their primary focus is on the logical structure of the scientific concepts. On the other end of the spectrum is naturalistic philosophy, which is closer to how scientists work and raises questions such as: What are scientific research goals? What methods are used to achieve them? How are the methods and results evaluated? As a result, the naturalistic perspective made clear that not all sciences are the same. It highlighted the distinct nature of the biological sciences and an urge for a different concept of laws due to the inability to find laws in the strict sense. Bechtel concludes that biologists and psychologists explain things differently, and that is by discovering and describing mechanisms.

3.1.2. Statistical Approaches and Their Limitations in Biological Understanding

About ten years after its initial presentation, the *DN* account began to receive a barrage of criticism with counter-examples based on two main points (Salmon, 1999). First, the existence of cases that, while meeting the model's requirements, failed to be good explanations, and second, the fact that bona fide explanations did not conform to the model's specifications. Particularly, regarding the second line of criticism, it has been observed that many good explanations draw from statistical rather than universal laws, that is, laws claiming that a certain fraction of *A*'s are *B* and not that all *A*'s are *B*. Especially for biomedical science phenomena, such as explaining the effectiveness of therapeutic regimens, the rule is the presence of statistical laws rather than universal generalizations. The introduction to the discussion of cases explained by statistical generalizations motivated Hempel to include in his explanatory account cases where the explanans does not necessarily lead to the explanandum but makes it more probable. Thus, about 15 years later, he

returns with a comprehensive account that includes the *Inductive Statistical (IS)* model of explanation, where particular occurrences are explained through an inductive argument. Here, a statistical regularity and some antecedent conditions comprise the explanans, and the explanandum is inductively inferred with a certain degree of probability. A more general definition given for the explanation is the following: "...an argument to the effect that the phenomenon to be explained ...was to be expected in virtue of certain explanatory facts." (Hempel, 1965, p. 336). This definition covers both the *IS* type of explanation and the *DN* type that has the same structure. Both models are arguments with premises that are either initial (antecedent) conditions or lawlike generalizations, and both have as a conclusion the phenomenon to be explained. However, in the *IS* model, the lawlike statements are not universal but statistical generalizations. The premises make the occurrence of the explanandum more probable.

The main target of the critique that the accounts of nomological explanation received was the fact that both of the offered models focus on the logical form of a scientific explanation. Salmon (1999) furthermore argues that the *IS* explanatory schema is, in essence, an incomplete *DN* explanation, so it is prone to the same problems identified in the latter. Why could someone suggest something like that? The reasoning for such a suggestion goes as follows. Accepting a deterministic position for everything in the world, we recognize the existence of causes that determine every event. The complete knowledge of these causes can predict the event's occurrence. The causes and laws governing the event's occurrence can be interpreted as the antecedent conditions in a *DN* argument, while the manifestation of the event constitutes the *DN* argument's conclusion. As noted in the previous section, there is symmetry between prediction and explanation in the *DN* model; thus, when only the antecedents are available, we can predict the occurrence of the phenomenon, while when we have the conclusion and we are looking for the antecedents, we are talking about explanation. However, in case we are missing some of the initial conditions and laws because science has not yet discovered their role, then the predictive power of the argument is diminished, and the conclusion follows with a probability, which means we have the form of an *IS* explanation. So, according to this argumentation, *IS* is indeed a *DN* explanation for which we do not (yet) know all the antecedents.

Moreover, the inductive-statistical type of explanation was attacked with various counter-examples (Salmon, 1999). It was stated that the high probability for the occurrence of a phenomenon is not necessary for a correct explanation and that it allows the participation of information in the explanans that is irrelevant to the explanation. The extensive critique that the Hempelian models

received goes on par with an epistemological non-reductive trend in sciences. Being in a reductionist stance, someone prefers to look for any possible explanation at the fundamental level of physical science where the *DN* and *IS* accounts could be more well suited. The opposite approach supports the explanatory significance of each science at its level. It requires examining traditional questions, such as those related to scientific explanation, in a new context. In biology, psychology, and neurosciences, this new context demands a total reconsideration in approaching what constitutes an explanation. Considering the methodologies of sciences such as biology, chemistry, psychology, and the like, the increased probability of some individual events is commonplace in the explanations we encounter, so the involvement of statistics in the explanatory endeavor is not the problem per se. Rather, it is how statistical regularities were used in the *IS* model. The problem is that it construes explanations as arguments. So, a little later, Wesley Salmon proposed his own Statistical Relevance (*SR*) model of explanation (Salmon, 1971).

Focusing on biological and biomedical phenomena, someone will probably conceive that they are very complicated and often correlated with many factors that contribute in different ways, increasing or decreasing the probability of an event occurring. Each factor shows a sort of statistical correlation with the explanandum. At the same time, there is a prior probability that the phenomenon under explanation will happen anyway, even without the mediation of these factors. Precisely, this is what Salmon (1971) wanted to capture by proposing the *SR* model of explanation. He suggested a comparison between the prior probability of the individual event's occurrence and the probability of its occurrence, including all individual probabilities of all the factors with an identified statistical correlation. According to one of Salmon's examples (Salmon, 1999), if we want to explain why a woman suffered a heart attack, we first investigate the factors that are positively or negatively related to a heart attack occurrence in general. Then, we construct an explanation that encompasses a) the prior probability of having a heart attack in the absence of the above factors, b) the role of each factor in the occurrence of a heart attack, and c) a list of the factors that apply to the woman in question (e.g., being a smoker, overweight, with high cholesterol and without regular exercising). Through this explanation, we can appreciate how much each factor contributes to the manifestation of the specific individual event. This model manages to capture how low probable events can be explained and incorporate only the statistically relevant factors.

3.1.3. Decomposition and Multilevel Analysis as Basis for Compositional Explanations

However, as Salmon himself noted, the statistical correlations in the account he proposed are not necessarily causal, and this is one of the problems that later led him to revise this model with another, which would better capture the causal role of factors that contribute to the occurrence of a phenomenon. Salmon suggested that a causal explanation should demonstrate how an event is situated in the causal nexus of the world (Salmon, 1984). He proposed a close association between the construction of causal explanations and the discovery of mechanisms that are often hidden, and the role of science is to reveal them. This approach is a revival of the causal-mechanical tradition. In contrast to the nomological explanations that focus on the logical form of an explanation, it gives special notice to the physical connection of the entities taking part in the systems that instantiate a phenomenon. In his Causal Mechanical (CM) model, he introduced the concept of a *causal process*, which is a physical process, and of *causal interaction*, which concerns how the space-time intersection of two causal processes results in the modification of the structure of both. In addition, he proposed five criteria to assess whether an explanation is causal. The criteria that Craver also used in his causal mechanical account are the following:

(E1) mere temporal sequences are not explanatory

(E2) causes explain effects and not vice versa

(E3) causally independent effects of common causes do not explain one another

(E4) causally irrelevant phenomena are not explanatory; and

(E5) causes need to make effects probable to explain them (Craver, 2007, p. 26).

Salmon's contribution, which emphasized the importance of the causal nature of processes, laid a solid foundation for compositional explanations. Another essential element for a new approach to explanations is the recognition that biological phenomena are explored at many levels of organization with part-whole relationships between them. Explanations of biological and psychological phenomena require recursive decomposition of the phenomena themselves or the systems that instantiate them, as these phenomena are process-related. According to the causal-mechanical tradition, a process can be explained by searching for processes and mechanisms at lower levels of organization. We have already seen this scheme of decomposition into lower-level parts and processes in Simon's (1962) work on hierarchical biological systems, in Kaufmann (1970) and Wimsatt's (1972) treatment of descriptive and interactional decomposition, and in Cummins'

(1975) proposal for the functional analysis of biological phenomena, an idea also applicable to psychology.

According to Cummins (2000), the explanatory questions in psychology and biology are of the type “How does it work?” rather than “What are the laws.” He argues that laws, as used in covering law models, are not explanatory since they tell us what happens but not why or how it happens. From this perspective, even the experimental testing of theories and hypotheses is a matter of confirmation, not explanation. Moreover, the ability to predict a phenomenon based on laws and antecedent conditions does not necessarily imply its explanation. Besides, the principles that generally govern nature do not apply in the phenomena of *special sciences* such as psychology; on the contrary, laws in those sciences are *laws in situ*, namely laws that concern special kinds of systems because of their particular constitution and organization¹⁸. In these cases, the explananda are effects or capacities, such as the ability to perceive depth or to be self-aware. Cummins (1983; 2000) argues that effects or capacities can be explained in terms of the systems' structure instantiating them, meaning their physical or functional constituents. This is where functional analysis comes into play, which is a form of systematic analysis of the system that has a capacity or of the capacity itself:

I call the application of property analysis to dispositions/capacities functional analysis. In the context of science, to ascribe a function to something is to ascribe a capacity to it that is singled out by its role in an analysis of some capacity of a containing system. When a capacity of a containing system is appropriately explained via analysis, the analyzing capacities emerge as functions. (Cummins, 1983, p. 28)

It's important to note that functional analysis, unlike system analysis, focuses solely on the capacities of a system rather than its material components. In a functionally analyzed system that exercises a specific capacity, the goal is to reveal its organized sub-capacities (or analyzing capacities). Since functional analysis examines only the capacity a system exercises rather than the entire system itself, this analysis can be performed without referencing any particular physical implementation. However, for a complete explanation, a more detailed account of the system's

¹⁸ Some sciences, like geology and psychology, which are not fundamental sciences, are called "special sciences". These phenomena can not be explained by general laws of nature. They need laws that govern special sorts of systems, called 'laws in situ' which are specifications of effects (Cummins, 2000)

implementation should be provided (Cummins, 2000). An example of Cummins's (1983) capacity analysis can be seen in the study of a biological organism. Its biologically significant capacities can be explained by analyzing various systems, such as the circulatory, digestive, and nervous systems. Each of these systems has its own capacities, which can be further broken down into sub-capacities. Cummins emphasizes the need for a complementary componential analysis of the system, as these sub-capacities are linked to specific components within it. However, it's important to note that Cummins considers functional analysis and componential analysis to be conceptually distinct.

As we have seen, the System Tradition has made a crucial contribution to the development of explanations based on decomposition. In response to disagreements over the applicability of explanatory models based on the logical structure of propositions corresponding to laws or antecedent conditions, Systems Biology has emerged with a strong reaction. This field offers a different perspective on biological phenomena, particularly through the application of research tools that employ mathematical, computational, and synthetic modeling strategies derived from the broader field of complexity science methodologies. The focus here is often on numerous parts and networks of interactions, with multiscale models being used to investigate phenomena across various levels of organization. This new trend in biological sciences has further emphasized the need for novel approaches to scientific explanation. Consequently, hybrid explanatory strategies have been proposed to accommodate the dynamic nature of related phenomena, combining more classic decompositional strategies with complex system analysis tools (Green, 2019). Clearly, the discourse has moved away from older deductive and inductive explanatory accounts and their associated arguments. Many now consider the decomposition of the system that is the target of explanation to be the common starting point for explaining biological phenomena.

Gillett (2016) argues that Cummins' proposal for functional analysis, the mechanists' proposal for mechanistic explanation, and even many reductive approaches seeking explanations at lower levels of science or nature are all compositional explanations exhibiting two essential features. The first, which he calls *Piercing Explanatory Power (PEP)*, refers to the ability of compositional explanations to explain entities of one kind at one level in terms of qualitatively distinct kinds of entities at another level. The second, termed *Ontologically Unifying Power (OUP)*, refers to the feature of a compositional explanation to demonstrate that the entities of the explanans and the explanandum, which occupy different levels, are, in some sense, the same. Carl Gillett has compiled the essential elements of compositional relations in a compositional explanation in an extensive presentation of what he calls *scientific composition*. This will be the subject of the next section and will serve as a

guide to understanding how explanations through multilevel decomposition work. The new-mechanistic explanatory approach, which will be presented in the following chapter, shares many common features with this account of compositional explanation.

3.2. GILLETT'S FRAMEWORK ON *SCIENTIFIC COMPOSITION*

In his recently published book “Reduction and Emergence in Science and Philosophy,” Carl Gillett (2016) offers a detailed framework for the notion of compositional explanations in science. The book aims to state his position in the debate mentioned in the title, and he uses this framework as a basis. His framework also provides an ontological and methodological starting point that can be used for a deeper understanding of scientific composition and decompositional methodologies. He develops the framework by examining how science works and how scientists produce explanations for various phenomena from the physical sciences to the biological domain.

As defined by Gillett, the term *compositional explanation* refers to “explanations of higher-level entities built around lower-level entities taken to compose them” (Gillett, 2016, p. 2) and addresses collective phenomena and their components. These phenomena are abundant in nature and reflect the common sense's image of a world occupied by entities such as individuals, properties, and processes that can be decomposed into smaller constituents. For example, a human body comprises systems with organs and particular kinds of tissues consisting of cells and fluids. Similarly, a pile of breakfast cereals is made up of tasty clusters consisting of pieces of oats and nuts, which in turn consist of starch, proteins, fat, and so on. We can apply such a decompositional analysis to most things that occupy our everyday lives. To explain the behaviors that these collections of entities exhibit, we need a qualitative analysis of the components and their behavior at lower levels.

A compositional explanation is closely related to what Gillett terms *everyday reductionism*, i.e., the pursuit of such explanations, and those have as a product the so-called *Scientifically Manifest Image*.¹⁹ The Scientifically Manifest Image consists of productive explanations that involve entities

¹⁹ This concept combines two other notions that Wilfrid Sellars (1963) first proposed: the *scientific image* and the *manifest image* of the world. According to the first, the world is presented as a product of scientific inquiry; for instance, science has revealed that a human is an organism consisting of body tissues, organs, and biomolecules. On the contrary, manifest image relates to an everyday perception of the world; It is not scientifically derived, rather it is a kind of personalistic knowledge.

at the same level combined with compositional explanations that posit relations between entities at different levels. According to Gillett, this image aligns with the idea of *non-reductive physicalism* in philosophy because it accepts the composition of higher-level entities from lower-level constituents while also acknowledging the determinative role of entities at their own level.

To understand Gillett's framework, it is necessary to highlight two essential starting points in his account. The first is a distinction between the determinative role of productive versus compositional relations and, thus, productive and compositional explanations. The second is his division of the entities involved in these relations into four ontological categories that are interconnected and interdependent: individuals, relations/properties, powers, and processes. In what follows, I will proceed to the first point.

Gillett distinguishes between *productive* and *causal* relations, noting that while all productive relations are causal, causation is a broader concept. Every productive relation is causal, and productive determination is a causal determination. Productive relations exist only between entities at the same level, a principle Gillett terms "*productive closure*." These are intra-level, horizontal relations. In an explanatory context, an entity's productive role refers to its relation with processes at the same level. Conversely, *compositional relations* are inter-level and vertical, involving lower-level entities that collectively and non-causally determine upper-level entities through constitution. Thus, when multiple lower-level entities' productive roles result in an upper-level entity's role, this relationship is compositional rather than productive or causal. Gillett argues that a comprehensive explanation must incorporate both productive and compositional aspects, effectively telling two perpendicular stories: one horizontal (productive) and one vertical (compositional).

In brief, the differences between productive and compositional relations, as they arise from Gillett's (2016) ideas are listed below:

PRODUCTIVE	COMPOSITIONAL
Horizontal	Vertical
Intra-level	Inter-level
Triggering/manifestation of powers	No triggering/manifestation of powers
Temporally extended	Synchronous
Between wholly distinct entities	(In some sense) same entities

Mediation of force	No mediation of force
Transfer of energy	No transfer of energy
Causal determination	Non-causal determination

The second key point in Gillett's framework is the ontological foundation of his position on scientific composition. This involves a breakdown into four ontological categories, known as the *Dimensioned view of realization* (Gillett, 2002). This view contrasts with the *Standard* or *Flat view* supported by philosophers such as Jaekwon Kim (1998) and Sydney Shoemaker (2001), which describes realization relations as involving only different properties rather than the four categories of entities that Gillett proposes²⁰.

As previously mentioned, composition in Gillett's framework is a relation between many lower-level entities that collectively compose one entity at an upper level. Gillett proposes four interrelated types of entities that are always found together: *individuals*, *properties/relations*, *powers*, and *processes*. All of these participate in compositional relations, resulting in four types of compositional relations. Consequently, lower-level powers, properties, individuals, and processes collectively *comprise*, *realize*, *constitute*, and *implement* their corresponding higher-level counterparts. Regarding the relations between these entities within the same level, Gillett presents the following schema:

1. Individuals are located in space-time.
2. Individuals instantiate properties and relations.
3. These properties/relations contribute powers to the individuals.
4. Based on these powerful properties/relations, individuals can participate in productive processes.

²⁰ Furthermore, in standard (or 'flat') view of the metaphysics of realization, realization cases are differentiated only based on one ontological dimension regarding different properties. Namely, the flat view of realization ascribes to any individual the same realized and realizer properties. Thus, a realized and a realizer property are both at the same level since they are properties of the same object or individual (Gillett, 2002).

It's important to note that individuals and properties/relations are individuated through their connection to a process. Gillett emphasizes that scientific composition concerns only these *powerful* properties and relations, which are responsible for individuals' productive powers. With such powerful properties and relations, an individual *bases* a productive process. This schema addresses the deeper question of why specific individuals base specific processes²¹.

Having presented these two foundational points, we can now delve into the core of Gillett's proposal on scientific composition. His idea is synthesized by two main elements: the notion of *working components* and the concept of *joint role-filling relations*. Gillett states that “scientific composition involves *working components* bearing *joint role-filling relations*” (2016, p. 81). He further elaborates: “I highlight how we have compositional relations in the sciences when we have collectives of entities with certain roles whose relations to each other allow them to jointly fill the qualitatively distinct productive role individuated of the relevant composed entity” (p. 12).

To better understand Gillett's proposition, two key concepts require clarification: (1) Working Components: These are entities belonging to any of the four ontological categories mentioned earlier (individuals, properties/relations, powers, and processes), individuated by their roles in associated processes. Importantly, not every entity related to the composed phenomenon (e.g., those with merely spatial containment relations or serving as background conditions) qualifies as a working component. To be considered a working component, an entity must actively participate in or have a role in a productive process. (2) Joint Role-Filling: This concept captures the collective nature of scientific composition. Under specific circumstances or background conditions, multiple entities with distinct roles *collectively constitute* the role of the composed entity at the upper level. Consequently, the function or “work” that the upper-level entity performs is essentially the combination of the works performed by the role-fillers at the lower level. Gillett emphasizes that this compositional relation is determinative, implying a natural necessitation: whenever the lower-

²¹ Gillett's account of the scientific composition offers not only a description of the construction of scientific explanations but also a metaphysical foundation of why individuals participate in specific processes and where do they base their causal (productive) powers. New-mechanistic explanatory approaches focus on individuals and processes only, e.g. MDC (2000) and Craver (2007) refer to organized entities and activities, while Bechtel (2008) refers to organized component parts and operations, leaving the metaphysical part out. Nevertheless, properties are implied in entities as they participate in certain processes based on specific properties they possess.

level role-fillers exist under the appropriate conditions, the upper-level composed entity necessarily exists as well.

Gillett (2016, p. 71) provides a comprehensive list of characteristic features of scientific composition, which completes the presentation of his account. Many of these features are implicitly mentioned or alluded to in the previous discussion. When combined with an account of relevant intra-level productive relations, these features can be used to construct a compositional explanation. In the next section, I will apply this compositional account to a biological case involving the entire human body.

To begin with, the relata in scientific composition are working entities, meaning their associated processes individuate them. Second, interlevel compositional relations are not productive, hence not causal; they are synchronous and apply between entities that are essentially the same. For instance, when considering a cell and its constituent intracellular structures, we cannot assume causal relations between the cell and an organelle because the organelle is part of the cell, and all organelles organized within the cytoplasm collectively form the entity we call a cell. Third, compositional relations are *mass-energy neutral*, meaning there is no mass or energy exchange between relata at different levels because they are the same entity, as mentioned earlier, with lower-level entities composing upper-level entities. Fourth, when there are more than two levels, compositional relations are transitive across all levels. This means that if lower-level entities compose middle-level entities, and middle-level entities compose upper-level entities, then lower-level entities also compose upper-level entities. Fifth, despite being transitive, scientific compositional relations are asymmetric; the synthetic capacity flows only in one direction: from lower to upper levels, not vice versa. Furthermore, they are not reflexive, meaning that in compositional explanations, scientists do not assume that individuals are parts of themselves nor that processes implement themselves.

The seventh feature is essential in Gillett's account because it underpins the notion of *qualitative emergence*, which concerns novel properties ascribed to higher-level individuals. Entities residing at different levels and compositionally related to each other are usually qualitatively distinct: upper-level entities possess powers and properties not held by the lower-level entities that jointly compose them. For example, a lipoprotein membrane is qualitatively significantly different from the individual proteins and lipids that form it. Eighth, scientific compositional relations are many-to-one, meaning that many individuals, properties, powers, or processes together compose the

corresponding upper-level entity. Ninth, under appropriate conditions, components naturally necessitate the composed entity, not due to the triggering or manifestation of powers as in intra-level productive determination, but because the composed entity is, in essence, all its components. The tenth common feature is analogous to the well-known multiple realization principle but broader, as it involves all four kinds of entities. Gillett uses the term '*multiple composition*,' which essentially states that different teams of component entities can combine to separately give rise to tokens of the same type of higher-level entity. Eleventh, any supposed composition will occur only if certain *background conditions* coexist. This implies that the components' existence and their interactions are not sufficient to produce the composed entity. However, an entity that jointly (with other entities) suffices for scientific composition is not necessarily a component, as it may be a background condition.

The last four features were previously mentioned in Gillett's ontological account of the role that different kinds of entities play in scientific composition and represent his dimensioned view of realization. He refers to them as the *coadunative* aspects of compositional relations and describes them as follows: "Scientists posit packages of compositional relations, in part-whole, realization, comprising, and implementation relations, between the individuals, properties, powers, and processes at different levels. Compositional relations between various categories are thus always found together, as our last four features together imply, and we thus always have *Dimensions* in scientific composition through its always involving relations amongst various categories of entity." (Gillett, 2016, p. 79) Thus, the twelfth common feature is that compositional relations always involve part-whole relations between individuals spatially contained within the individual they collectively constitute. Thirteenth, composition always involves the collective realization of properties; fourteenth, it involves the comprising of powers; and finally, fifteenth, it always involves lower-level processes that together implement, or potentially implement higher-level processes.

In his (2016, p. 80) work, Gillett lists the essence of scientific composition into fifteen core features. These characteristics offer a comprehensive snapshot of how entities at different levels interrelate in scientific explanations and form the backbone of Gillett's compositional framework:

1. Relata are working entities
2. Non-productive determination relations
3. Mass-energy neutral relations
4. Transitive

5. Asymmetric
6. Irreflexive
7. Qualitatively different relata
8. Many-one relations
9. Natural necessitation for the composed entity
10. Multiple composition is allowed
11. Hold only under background conditions
12. Always involve part-whole relations with spatially overlapped individuals
13. Always involve comprising powers
14. Always involve realizing properties
15. Always involve implementing (or the potential of implementing) processes.

3.3. ELBOW MOVEMENT AS A MODEL FOR MULTILEVEL BIOLOGICAL ANALYSIS

So far, in the first chapter, I introduced the concepts of hierarchy and levels, focusing on biological phenomena. The second chapter explored the ideas of composition and decomposition in component functions or parts, examining their relations both generally and within complex biological systems. I also discussed organizational principles governing the formation of hierarchies, providing essential insights into potential analytical and explanatory strategies. In this chapter, I have reviewed the challenges of applying traditional nomological and statistical explanatory methodologies to biological and psychological fields. I then traced the evolution of ideas and arguments leading to the more appropriate compositional explanatory schema. Subsequently, I presented Carl Gillett's account of *scientific composition*, which offers a detailed framework for constructing scientific explanations in systems that can be analyzed across levels with compositional relations. To further examine and illustrate these concepts and ideas, I will now turn to a biological case study: elbow movement. This example will serve as a practical model for multilevel biological analysis, allowing us to apply the theoretical foundations established in the preceding chapters.

In what follows, I employ a compositional explanation in a biological case that can be analyzed and decomposed across multiple levels. The chosen biological case revolves around the scenario where an individual employs muscular force to flex her forearm toward her upper arm. This specific case was selected deliberately for its potential to provide a comprehensive analysis spanning multiple

levels of organization. The analysis begins at a higher level, encompassing the entire body's interaction with the environment, and progresses through several levels, ultimately reaching the subcellular and molecular levels, which represent the lowest level of this analysis. Analyzing in this way the elbow movement, offers a valuable opportunity to examine the concept of multiple levels and how explanations unfold across them.

Another compelling reason for selecting this particular biological example is its suitability for detailed decomposition from both material and functional perspectives. The system is decomposable to varying degrees, although it's important to note that this characteristic is not universal across all biological systems. Some systems are more decomposable, while others are not. This particular case combines numerous features that make it an ideal candidate for applying compositional explanations in biological contexts, as we will observe when systematically breaking it down to elucidate the underlying phenomena. Decomposability, however, does not imply that the diversity and organization of components and functions at each level are simple and easily detectable. As we will see, these aspects vary depending on the level of organization where we focus our analysis at any given time, highlighting the complexity inherent in biological systems even when they are amenable to multilevel analysis.

To begin, I will briefly outline the essential elements necessary for understanding the kinetic mechanisms responsible for this specific movement. A systematic scientific study of these mechanisms requires decomposing the system and its functions at lower levels, which implies a higher resolution of analysis. At each different level, distinct yet complementary narratives emerge. The comprehensive explanation is a synthesis of all intra-level (productive) and inter-level (compositional) relations. This ability to examine the phenomenon in two directions simultaneously - the horizontal productive and the vertical compositional - is a key strength of compositional explanations. After exploring how levels of explanation emerge in the system under examination, in the final section, I will focus on a particular detail of the mechanism: the mechanism of muscle contraction. By exploring specific aspects of this mechanism, I aim to illustrate more comprehensively the concepts of productive and compositional relations. This approach will also provide a practical demonstration of how decomposition leads to explanation.

3.3.1. Anatomical and Physiological Essentials of Forearm Flexion

The human body interacts with the environment by receiving information through its perceptual systems. This information is processed and combined with prior knowledge and experience to

produce behaviors that constitute our behavioral repertoire in the world. The body behaves as a whole, existing in a continuous feedback cycle that connects the internal with the external. It employs its systems in a constant, interactive, and coordinated manner. Each system plays a distinct role in regulating human behavior. In the specific case of body movement, the focus is on three systems that play a leading role. These are the nervous system, responsible for commanding and controlling movement; the muscular system, responsible for producing and exerting force; and the skeletal system, which serves as the body's frame upon which muscular force is applied, resulting in changes in position and joint movement. Of course, other bodily systems also play essential roles. The circulatory system provides muscles with oxygen and nutrient-rich blood, the respiratory system ensures a constant supply of oxygen, and the endocrine system regulates glucose supply to muscle cells. Their role is supportive, acting as background and enabling conditions for the smooth operation of the three main systems, which together comprise the human motor system.

The motor command and orchestration of movement occur in the brain, involving multiple central structures. The motor cerebral cortex generates motor commands, while the brainstem, cerebellum, and basal ganglia are responsible for control, coordination, and movement adjustments. Several spinal cord structures also play a role, connecting the central processing areas with peripheral nerves. The peripheral part of the process includes sensory neurons connected to muscle receptors and motor neurons, which continuously cooperate dynamically through feedback loops to adjust movement. Motor neurons enter muscles and convey the motor pattern generated in the brain through the *motor plates*, which are neuromuscular synapses. Each motor neuron innervates many muscle cells (muscle fibers), and this entire unit is called a *motor unit*.

In the case of elbow movement, the elbow joint connects the distal humerus with the proximal ulna and radius, which together comprise the forearm. Multiple muscles attach to the forearm and move it, serving various roles such as bending the elbow, stabilizing the joint, and mitigating or neutralizing unwanted movements. The primary elbow flexors are the biceps brachii, brachialis, and brachioradialis, with the biceps brachii being the strongest and most essential for executing this specific movement²². The motor command received by the musculocutaneous and radial

²² For a more detailed presentation of the muscles, bones and elastic elements that constitute the elbow joint consider “Kinesiology: the mechanics and pathomechanics of human movement” (Oatis, 2009)

nerves²³, initiates a contraction process that reduces the muscle's length. This contraction exerts a traction force on the two distal ends of the muscle, the tendons. Tendons, which are passive elements of the mechanism, transfer the traction force to the connected bones, pulling the more distal part of the joint upwards into a bending position.

What happens inside the muscle concerning its constituent parts and processes? When the brain sends an electrical stimulation pattern through the spinal cord to the motor units, the stimulation spreads to the muscle fibers, triggering calcium release through the *sarcolemma* (the muscle fiber's cell membrane). Each muscle fiber comprises numerous myofibrils, which in turn consist of a long series of sarcomeres connected at specific points called *Z-lines*. Sarcomeres are composed of actin and myosin filaments which are partially overlapping. Calcium plays a crucial role in initiating muscle contraction by activating a molecular mechanism within the sarcomeres called the *cross-bridge cycle*²⁴. The key players in this cycle are two proteins, actin and myosin, which are folded into the long filaments that constitute the sarcomeres. Two accessory proteins, troponin and tropomyosin, bound to the filaments, regulate the process. Additionally, an abundance of ATP molecules provides the necessary energy for the entire procedure. Cross-bridge cycles form continuously throughout the contraction, causing a sliding movement between the partially overlapping actin and myosin filaments. This mechanism is described by the *sliding filament theory*²⁵. The sliding of filaments leads to sarcomeres' contraction. Sarcomeres, therefore, are the functional units of muscle contraction.

²³ For the forearm flexion two nerves are responsible: musculocutaneous nerve which innervates biceps brachii and brachialis and radial nerve which innervates brachioradialis (Floyd & Thompson, 2006)

²⁴ The model of *Cross-Bridge Action* was described in two articles by Huxley and Simmons. The mathematical formulation of the idea was described in "Proposed mechanism of force generation in striated muscle" (Huxley & Simmons, 1971) and the qualitative description in "Mechanical transients and the origin of muscular force" (Huxley & Simmons, 1973).

²⁵ The "Sliding Filament Theory" for muscle contraction is currently the most prevalent theory that explains the mechanism of muscle contraction. It started as a theory in 1954 with two papers published in Nature under the general title "Structural changes in muscle during contraction." The first from Andrew F. Huxley and Dr. Rolf Niedergerke, entitled "Interference microscopy of living muscle fibers", and the second from Dr. Hugh Huxley and Dr. Jean Hanson, entitled "Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation". These ideas were further enriched with the '*kinetic model of contraction*' which was developed in 1957 by A.F. Huxley and dealt with how force relates to velocity and the mechanism of energy release during isometric and shortening contractions (Rall, 2014).

As each muscle fiber contains many myofibrils composed of many sarcomeres serially connected, the collective reduction in sarcomeres lengths results in a reduction of the overall muscle fiber length. Multiple muscle fibers connected to a motor neuron at a motor plate form a motor unit. The contraction of several motor units results in an observable contraction of the muscle and, consequently, a joint movement. During muscle contraction, the signal generated in the brain and conveyed by the motor neuron triggers a higher-level mechanism called *motor unit recruitment*. This mechanism describes how the muscle is activated during movement execution. Factors such as the load to be overcome (e.g., when lifting a weight) and the desired speed dynamically influence the number and type of motor units recruited.

These mechanisms for muscle contraction are nested, meaning there is a containment relation between higher and lower mechanisms. For instance, at the highest level, the muscle is activated, and its power is adjusted through motor unit recruitment. At an intermediate level of organization, muscle fibers reduce their length according to the sliding filament mechanism. At the lowest level, filaments slide on each other via the formation of cross-bridge cycles. In the following sections, we will explore these mechanisms and their role in a compositional explanation in more detail.

3.3.2. Decomposing the Elbow Movement System to Reveal Hierarchical Organization

To comprehend the compositional explanation approach, it is essential to understand the concepts of levels, hierarchy, and decomposition and how they apply to an explanation, let us say, to the elbow flexion example. In an explanation, levels emerge as we repeatedly inquire about *how* something works or *why* an observed phenomenon occurs. This reflects a deepening of our understanding and is expressed through more detailed descriptions with higher resolution. As we descend through levels of biological organization, we encounter smaller entities and processes that occur over shorter durations. Levels of biological organization also arise when we ask *what* a certain entity consists of, a question applicable to both individuals and functions. Hence, as we have explored in the first chapter, levels of organization are defined not only by size or duration but also by part-whole relations between lower-level and upper-level functions and material components. The same stands for levels recruited in compositional explanations since they are based on a hierarchical conception of the phenomena and systems to be explained.

Part-whole relations, size of entities, and duration of processes play an important role, but also very crucial is the consideration that when levels are used for explanatory purposes, their application has a local character, as has been strongly emphasized by Craver (2007) in his account of levels of

mechanisms in neuroscience. An implication of this local character is that cells from different tissues in the body are not necessarily considered to be at the same level, even if they are of similar size, because they concern different explanations, i.e., explanations for phenomena occurring in different systems. For instance, in this context, a random tissue cell is not at the same level as a muscle cell since it relates to a different bodily function. The criteria for compositional levels are quite restrictive, focusing on functional relations and hierarchical organization rather than mere physical characteristics.

To answer both types of questions – *how something works* and *what it consists of* - requires the system's functional or structural decomposition. Levels emerge through recursive decomposition, but they also serve as a reference point to guide further decomposition. The recursive application of decomposition results in a hierarchy of nested levels occupied by entities and processes. The summarized description of the forearm flexion presented in the previous section mentioned entities and processes belonging to different levels of organization, ranging from the bodily behavior of movement perceived by simple observation at the higher level to the molecular processes of muscle contraction revealed through scientific research at the lowest level. To understand the explanation in a more systematic way, we will attempt to describe the mentioned mechanisms in a hierarchical manner, where every behavior described at a specific level of organization will be decomposed into component parts and processes that compose and explain it at a lower level.

At the highest level, concerning the entire body where forearm flexion can be physically observed and described, two body parts - the arm and the forearm - engage in a bending movement. In other words, breaking down the body's behavior at this *whole body level* results in describing two distinct body parts. The next decomposition at the *body parts level* reveals the involvement of three different types of tissues that constitute each body part: nervous, skeletal, and muscular. The cooperation of these three tissue types involves the generation of a command from the nervous system targeting the muscles that pull the forearm bone. This level of *body organs* consists of the specific muscles that collectively flex the forearm (biceps brachii, brachialis, and brachioradialis), the nerves that innervate them (musculocutaneous and radial), and the bones of the joint (humerus, ulna, and radius) along with their periarticular elements.

At this stage of the decompositional analysis, a bifurcation emerges concerning the subsystems that we examine. More specifically, this level, which consists of muscles, nerves, bones, and joints, is the

lowest level where all three tissue types (skeletal, neural, and muscular) that cooperate in the flexion mechanism interact. From this point onwards, the motor system splits into two distinct subsystems with operations that no longer interact: the skeletal system (bones and joints) and the neuromuscular system. Any other decomposition of bones and articular elements will yield levels functionally and structurally independent of those involving nerves and muscles. Regarding the bones and joints of the elbow, the story goes as follows: They convert muscle tension into motion, and this can be explained by the laws of biomechanics. The joint is studied as a simple machine - a lever consisting of bones connected at the elbow joint - with muscular tension representing the force that pulls the bones against resistance (usually an external load or the forearm's weight). The lever's fulcrum, or axis of rotation, is located centrally at the elbow joint (Floyd & Thompson, 2006). It is neither meaningful nor practical to continue describing the entire tree graph that results from all the recursive decompositions we apply to examine elbow movement. We will focus on a subunit of the overall mechanism in order to make the whole process understandable, that is, to demonstrate the depth of analysis that can be achieved through repeated application of decomposition for explanatory purposes. Consequently, we will examine only nerves and muscles hereafter, which continue to interact as our decompositional analysis deepens.

As we delve deeper into the explanatory description, we find that muscles and nerves, being functionally interconnected, form a subsystem that serves as the active component of the movement mechanism, while bones and joints constitute the passive component. Within this active component, there is a strong interplay between elements of the nervous system and those of the muscular system. Exploring the *neuromuscular level*, we find that several muscle fibers are connected to a single motor neuron, constituting a *motor unit*. At a lower level, each muscle fiber belonging to the same motor unit connects to a nerve fiber ending at a specific point called *motor plate*. By further decomposing the motor plates (at the *motor plates level*), we focus on muscle fibers (at the *muscle fibers level*) comprised of *myofibrils*. Here, we enter the realm of muscle tissue, leaving the nervous tissue behind, as we did with bones and joints at the former, higher levels. At these lower levels, where descriptions primarily focus on muscle tissue, the nervous system's involvement is restricted to triggering the release of calcium ions (Davis, Tikunova, & Janssen, 2019; Oatis, 2009), which signals the initiation of muscle contraction. Our aim henceforth is to examine how a muscle fiber's ability to generate tension at its ends is explained. This compositional explanation will be presented in the next section.

When we consider the questions posed by scientists in the field to explain body movements, along with the entities and processes they describe in their explanations, a hierarchy of eleven levels emerges. In what follows, a representative hierarchy of explanatory levels is presented, with the first level (referring to the whole body) being the highest and the last level (referring to the molecular mechanism of muscle contraction) being the lowest. For each level, I list the lower-level entities that compose it and the primary process that is executed. It's important to note that these levels have a local character; they are closely related to the specific phenomenon we aim to explain. This is also the reason why specific names are used for their designation, such as 'level of motor plates,' 'level of myofibrils', and the like. As we descend to lower levels, we'll observe that they become increasingly specific and narrowly focused²⁶. This progressive narrowing of focus reflects the increasing depth and specialization of our understanding as we delve deeper into the mechanisms underlying forearm flexion.

- 1) Level of the *whole body*: composed of body parts (e.g., arm, forearm). Process: forearm flexion.
- 2) Level of *body parts*: composed of different types of tissues that comprise various organs (e.g., bones, muscles, nerves). Process: Cooperation for control and execution of the movement.
- 3) Level of *body organs*: composed of tokens of organs such as specific muscles, nerves, and bones (e.g., humerus, ulna, radius, biceps brachii, brachialis, brachioradialis, musculocutaneous nerve, and radial nerve). Process: biomechanics of movement.
- 4) *Neuromuscular level*: composed of motor units. Process: Muscle power generation
- 5) *Level of motor units*: composed of bundles of muscle fibers of three different types (slow-twitch oxidative *red fibers*, fast-twitch oxidative/glycolytic *intermediate fibers*, and fast-twitch glycolytic *white fibers*) and a motor neuron. Process: power and speed control (motor unit recruitment)
- 6) *Level of motor plates*: composed of a motor neuron branch connected to a single muscle fiber. Process: Signal transmission
- 7) *Level of muscle fibers*: composed of myofibrils. Process: production of muscle fiber contraction
- 8) *Level of myofibrils*: composed of sarcomeres. Process: Shortening of myofibrils
- 9) *Level of sarcomeres*: composed of actin and myosin filaments. Process: Z- lines convergence

²⁶ The presented decomposition reveals various anatomical and physiological levels as they emerge in descriptions found in “Functional Morphology of the Striated Muscle” (Kilarski, 2019), “Motor Units and Muscle Receptors” (Celichowski & Krutki, 2019), “Muscle Function: Strength, Speed, and Fatigability” (Enoka & Duchateau, 2019) and “Kinesiology: the mechanics and pathomechanics of human movement” (Oatis, 2009).

10) *Level of myofilaments*: composed of actin, myosin, troponin, tropomyosin, ATP, and Ca. Process: Myofilaments overlapping (*sliding filament theory*)

11) *Molecular level*: composed of structures within molecules. Process: *Cross-bridge cycle* formation

Upon our navigation through the hierarchical decomposition of the elbow movement system, we encountered significant bifurcations that shaped our explanatory focus. These bifurcations, such as the divergence of the skeletal and neuromuscular systems, guide our attention to specific branches of the hierarchy that are most pertinent to understanding forearm flexion. For instance, to comprehend the mechanism of forearm flexion, it is crucial to explain at a higher level of description how muscle force is transmitted. However, delving into the microstructural organization of osteocytes within the skeletal system becomes superfluous, as it does not directly contribute to our understanding of muscle flexion mechanisms. Similarly, while explaining how nerve signals propagate to muscle fibers at the motor plate level is essential for grasping muscle contraction, further decomposition of neural elements to analyze action potential creation in neurons becomes unnecessary once we've explained the neuron-muscle fiber connection.

This selective focus on the most relevant factors allows for a more targeted and comprehensive explanation. The decomposition process is thus guided by the specific phenomenon we aim to understand rather than by an indiscriminate breakdown of all system components. It's worth noting that levels of decomposition do not always align perfectly with explanation levels. The interaction profile between the system's parts and functions serves as an additional guide in this explanatory delimitation, offering valuable insights into which aspects warrant deeper exploration. The resulting hierarchy of levels functions akin to a ladder. As we descend, we gain an increasingly detailed and profound understanding of how the forearm flexes towards the arm. This approach reveals that an explanation is not a comprehensive pyramid of all possible compositional relations. Instead, it represents a specific path through this pyramid, defined by the questions we seek to answer in our quest to understand a particular phenomenon.

Initially, our guiding question was broad: "How does the forearm flex towards the arm?" As we progressed with decompositions, this question evolved to focus on muscle power generation: "How is the power that flexes the forearm generated?" Following the next decompositions, our inquiry further narrowed to the specific mechanism of muscle contraction: "How is the muscle contracted?" This approach to explanation allows us to navigate the complex interrelations between different levels of biological organization effectively. It provides a structured framework for understanding

how macroscopic behaviors, such as forearm flexion, emerge from microscopic processes at the cellular and molecular levels. By following this path through the explanatory pyramid, we can achieve a comprehensive understanding of the phenomenon while maintaining a sharp focus on the most explanatory relevant aspects at each level of description.

3.3.3. Muscle Contraction Mechanisms Through the Lens of Compositional Explanation

This section examines the muscle tissue mechanisms underlying the process of muscle contraction. We will explore these mechanisms across multiple levels of organization, from muscle fibers down to molecular interactions. Our analysis will demonstrate how a typical scientific description of muscle contraction aligns with Gillett's (2016) model of compositional explanation, which unfolds across multiple levels²⁷. Additionally, we will illustrate how the gradual decomposition of the system into progressively lower compositional levels contributes to a comprehensive explanation of the phenomenon. To achieve this, I will conduct a structural decomposition of a muscle fiber²⁸, revealing in detail the entities (or component parts, which Gillett refers to as “individuals”) involved in the processes at each level. This decomposition will demonstrate how individuals at one level collectively constitute the individuals at the level above, adhering to Gillett's criteria for compositional relations between individuals. *Figure 6* provides a visual representation of a muscle's structural organization, illustrating the inter-level compositional relations among the entities that constitute the muscle.

²⁷ In this presentation, I will narrow down my inquiry to the individuals and the processes, and I will leave out the part of the metaphysical foundation of the explanation through the properties/relations and the powers of the individuals because it does not fall within the objectives of the present work.

²⁸ An even more detailed presentation of the inner structure of the muscle fiber can be found in Kilarski (2019) and Oatis (2009, pp. 45-68)

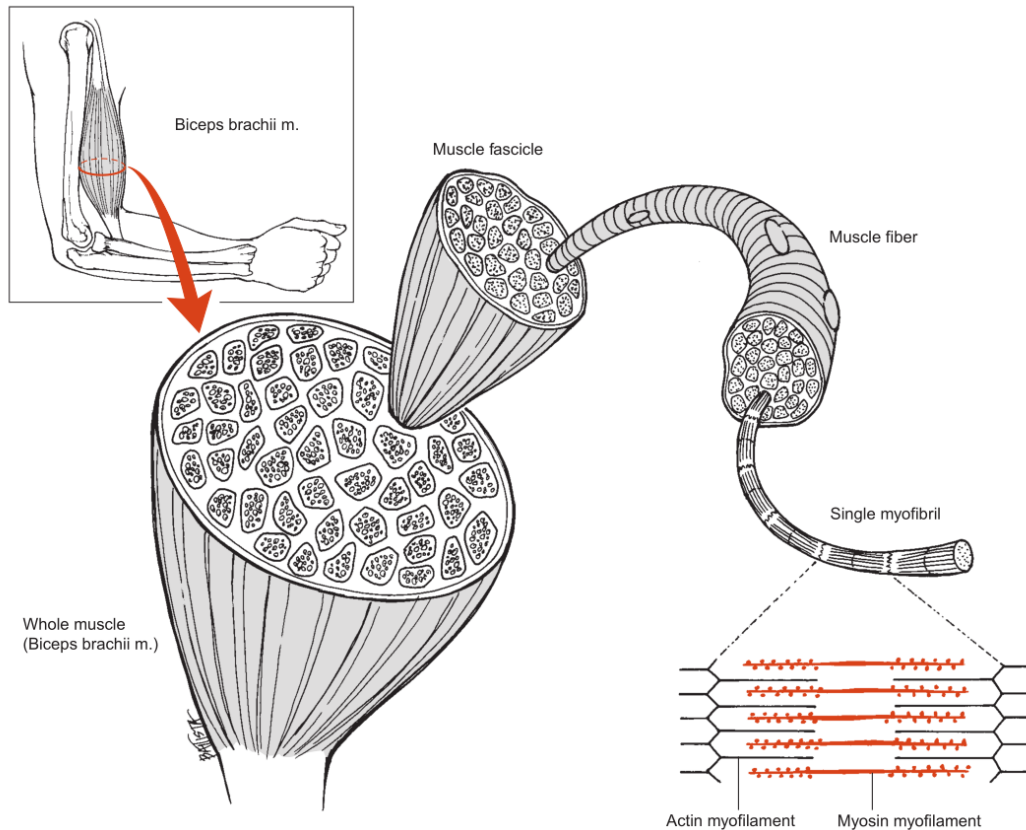


Figure 6: Organization of muscle. A progressively magnified view of a whole muscle demonstrates the organization of the filaments composing the muscle. Reproduced from Oatis (2009, p. 46)

A muscle's ability to exert traction on a bone, thereby moving it, stems from the collective contraction of its fibers. Coordinated by the nervous system, these fibers shorten and generate tension, which is then transferred to the bones via tendons. The production of this contractile force occurs within the muscle fibers or muscle cells. Each muscle fiber contains contractile structures capable of generating tension, as well as connective tissue that transmits this tension to the bone²⁹. The tension is collectively produced inside the muscle fibers, which are elongated cylindrical cells enclosed in the *sarcolemma* (the muscle fiber's cell membrane). These fibers contain myofibrils, which are parallel-organized structures responsible for contraction, numerous mitochondria strategically distributed to meet metabolic demands, and a network of membranes crucial to the

²⁹ The connective tissue, found between the contractile structures of the muscle fibers and forming the tendons that connect muscles to the bones, consists mainly of collagen and elastin fibers and constitute the passive restraint mechanism that cooperates with the active contracting mechanism of the muscle.

contraction process, which is called *sarcoplasmic reticulum*. The sarcoplasmic reticulum's primary function is to accumulate and release calcium ions, which play a vital role in the contraction mechanism. By storing and releasing these ions at precise moments, the sarcoplasmic reticulum helps regulate and facilitate muscle contraction.

The long, cylindrical myofibrils contained in the muscle fibers and encapsulated in the sarcoplasmic reticulum consist of a series of identical cylindrical segments called sarcomeres, which are the fundamental units of muscle contraction. A sarcomere is defined as the region between two *Z-lines* (also called *telofragms*) and contains two types of protein filaments: thick myosin filaments and thin actin filaments. The myosin filaments are longer and anchored at the center of the sarcomere, while the actin filaments are attached to the Z-lines on either end. Each actin filament spans two adjacent sarcomeres, with its free end extending towards the center of the sarcomere, where it overlaps with the myosin filaments. The thin actin filaments are composed of three proteins: actin, troponin, and tropomyosin. Tropomyosin molecules twist around each other

When an action potential is triggered at the neuromuscular junction, calcium ions bind to troponin, exposing actin's binding sites and initiating the contraction process. The thick myosin filaments comprise hundreds of rod-shaped myosin molecules, each consisting of three parts: a head, neck, and tail. The head contains binding sites for both actin and ATP (the energy source for contraction). The neck is a long, alpha-helical structure that acts as a lever arm during the cross-bridge cycle. The tail intertwines with other myosin tails to form a long double helix. Actin, being thinner than myosin, provides protein binding surfaces for the myosin heads. This intricate arrangement of protein filaments allows for the sliding filament mechanism of muscle contraction, where myosin heads attach to and pull on actin filaments, causing the sarcomere to shorten and the muscle to contract.

The entities (*individuals*) described above are the key players in the structural hierarchy of muscle composition, each playing a crucial role in the contractile process: They are *working entities individuated* by the contractile *processes in which they take part*. These contractile components³⁰ are related as *parts of a whole* in *many-one* compositional relations, forming a hierarchical structure where multiple lower-level individuals *jointly compose* a single higher-level individual. At the apex

³⁰ Many other entities, such as mitochondria and the sarcoplasmic reticulum, are not contractile components of the mechanism but work as enabling or background conditions.

of this hierarchy are muscle fibers, which are composed of numerous myofibrils. Descending the structural ladder, myofibrils are made up of multiple sarcomeres, which in turn consist of actin and myosin filaments. At the lowest level, these filaments are comprised of individual actin, myosin, troponin, and tropomyosin proteins.

It's important to note that these inter-level relationships are *not productive*, meaning that the higher-level structures do not generate the lower-level components. Instead, the individuals at different levels occupy the same physical space, they are *spatially overlapped*. For example, a series of sarcomeres occupies the same space as a myofibril because, in essence, they are one and the same structure viewed at different levels of organization. This spatial overlap applies to all compositional relations in the hierarchy. This structural arrangement aligns with the *synchronous* nature of muscle contraction processes. For instance, the sliding of filaments occurs simultaneously with the contraction of the sarcomere they constitute.

These compositional relations exhibit several other key properties mentioned in Gillet's (2016) compositional explanations' framework: (1) *Asymmetry*, viz., while multiple sarcomeres compose a myofibril, the reverse is not true - multiple myofibrils do not compose a sarcomere. (2) *Transitivity*: A sarcomere can be described as consisting of actin and myosin filaments (the next lower level) or as consisting of actin, myosin, troponin, and tropomyosin proteins (the bottom level). Lastly, (3) *non-reflexivity*: An individual is not part of itself; for example, a sarcomere is not a component of itself.

As we delve deeper into muscle contraction processes, additional features characteristic of compositional explanations will become apparent, further illuminating the complex interplay between these structural components. We begin by seeking an explanation for the phenomenon at the highest of the five levels under consideration (muscle fibers, myofibrils, sarcomeres, myofilaments, and molecular level). At the muscle fiber level, our explanandum is the production of muscle fiber contraction, which can be expressed through causal relations. The *intra-level productive explanation* at this level can be described as follows: when an electrical stimulus reaches the sarcolemma, it initiates muscle fiber contraction, resulting in (or causing) the exertion of a traction force at its ends. However, this raises two further questions: What does *muscle fiber contraction* entail, and how is it accomplished? Furthermore, *how is the traction force generated*? Both questions can be addressed by examining mechanisms performed by lower-level individuals and processes that compose the individuals and processes at this level.

Indeed, at the myofibril level immediately below, we observe a shortening in the length of parallel-organized myofibrils, which *jointly produce* a reduction in the muscle fiber's length. This mechanism is relatively simple, as its constituent entities are identical and lack complex organization. The length reduction of the muscle fiber corresponds directly to the length reduction of its components, namely the numerous parallel-organized myofibrils. Similarly, the contraction force produced by the muscle fiber is the sum of the individual contraction forces of its myofibrils, generated as a result of their shortening. These many-one relations jointly compose the explanation for the upper-level phenomenon: the muscle fiber is contracted because the length of its myofibrils shortens, and it exerts a traction force as a result of the sum of the forces produced by its myofibrils.

However, this explanation alone is again insufficient. New questions emerge. For instance, why does a myofibril shorten? To understand why a myofibril shortens, we must descend to an even lower level. Here, we discover that myofibril shortening results from the Z-lines of the sarcomeres moving closer together. This means that each sarcomere shortens, and the total shortening of the myofibril equals the algebraic sum of the shortenings of the serially arranged sarcomeres that compose it from one end to the other. In these examples, we observe the critical role of entity organization in producing the resulting upper-level phenomenon, which remains relatively simple.

As we continue descending to lower levels, we encounter increasingly complex mechanisms (Davis, Tikunova, & Janssen, 2019; Oatis, 2009). The sarcomere's shortening is explained at the next lower level by the sliding filament theory. According to this, the actin filaments that partially overlap with myosin filaments slide across myosin filaments towards the sarcomere's center. This sliding motion results in a reduction of the sarcomere's total length (see *Figure 7*). Here things start to get more complicated, and there is a feeling that we are getting closer to the answer, namely, how can a muscle fiber contract. Nevertheless, the explanation cannot be completed without describing the mechanism that takes place at the bottom level. There, through the description of the molecular processes, we can finally understand how such a force is produced that causes the filaments to slide, the sarcomeres to shorten, the myofibrils and the muscle fiber to contract, and finally to exert force on its ends at the tendon. This is explained through the process of cross-bridge formation.

Cross-bridges are energy-consuming connections between myosin heads and actin's binding surfaces. In this process, myosin necks pull against actin in a manner analogous to rowers pulling oars to propel a boat through water. The mechanism is initiated by calcium ions, which bind to

troponin, thereby enabling actin's binding sites to interact with myosin heads. Concurrently, after consuming ATP energy, myosin necks position themselves at an acute angle towards the sarcomere's edges. When the binding with actin becomes sufficiently strong, the strain on the myosin's neck is released, causing myosin to pull on actin. This generates a force directed towards the center of the sarcomere. The energy stored in the myosin necks is converted into kinetic energy, producing the sliding motion of the filaments. These processes are described in detail by the *Walk-Along Theory of Contraction* (see Figure 8), which is the most prevalent hypothesis that integrates all available scientific data into a mechanism explaining force production within sarcomeres (Guyton & Hall, 2006). This molecular-level description represents the foundational layer of the explanation, arguably completing our understanding of the muscle contraction process across all levels.

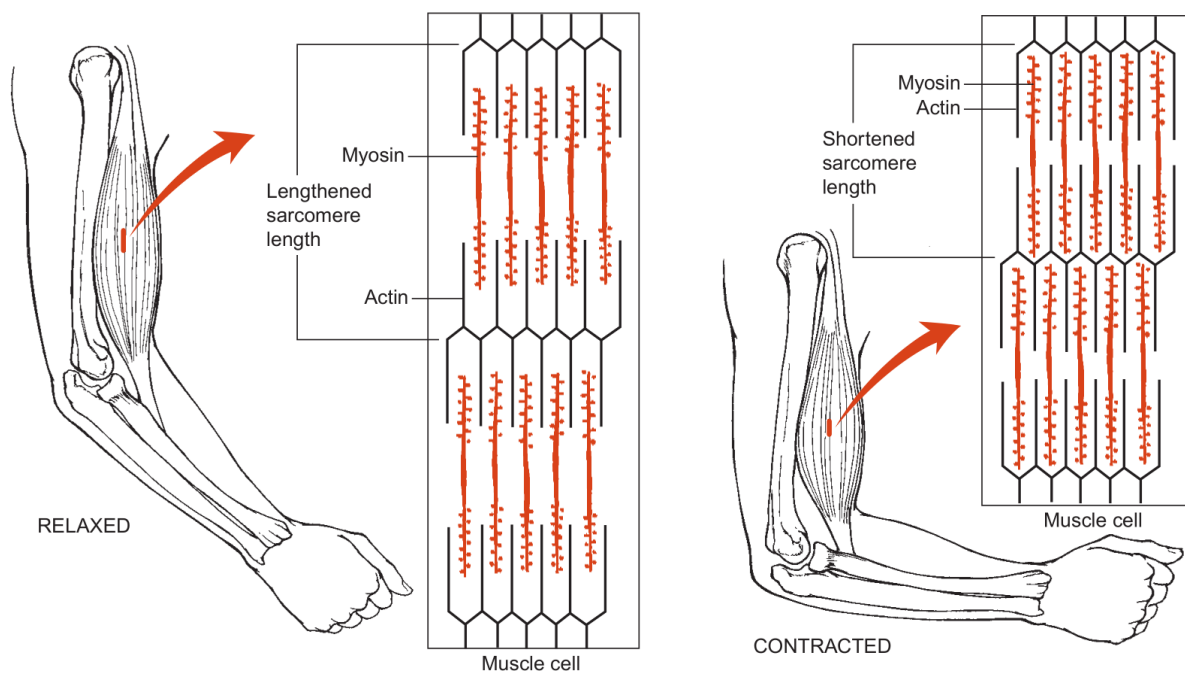


Figure 7: The sliding filament model. Contraction of skeletal muscle results from the sliding of the actin chains on the myosin chains. Reproduced from Oatis (2009, p. 47)

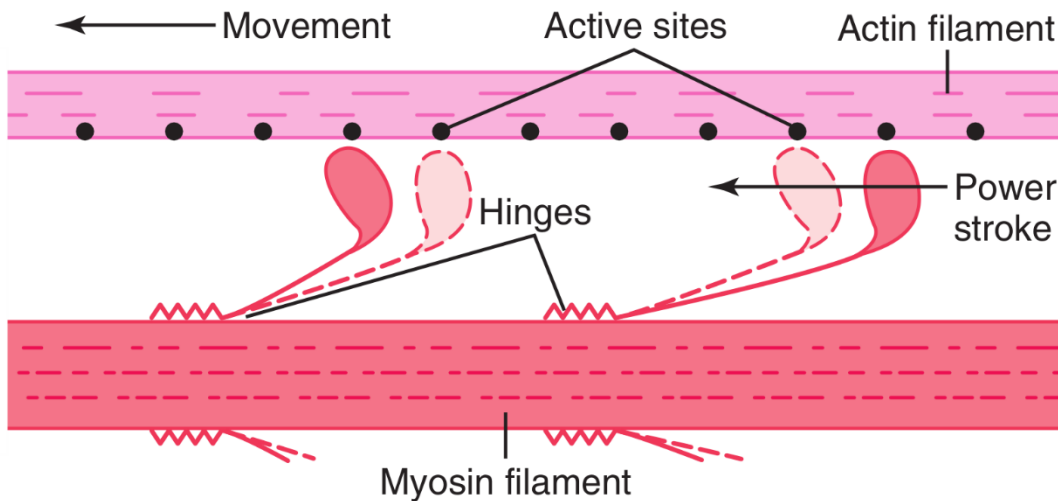


Figure 8: The "walk-along" mechanism for contraction of the muscle. Reproduced from Guyton and Hall (2006, p. 80). The walk-along mechanism for myofilament contraction involves the interaction between myosin cross-bridges and activated actin filaments. In this process, the heads of myosin cross-bridges attach to active sites on the actin filament, undergo a power stroke by tilting towards the arm, and then detach to repeat the cycle with a new active site further along the actin filament. This repeated cycle of attachment, tilting, and detachment causes the myosin heads to "walk" along the actin filament, pulling the ends of actin filaments towards the center of the myosin filament and resulting in muscle contraction.

This multi-level description of muscle contraction exemplifies Gillett's framework for scientific composition. At the molecular level, the productive processes of cross-bridge formation between actin and myosin proteins (the individuals) *together implement* the sliding filament process. Moving up, these lower-level productive processes involving actin and myosin filaments (individuals) *together implement* the process of sarcomere shortening at the next level. This pattern of implementation and composition repeats at each successive level, from the molecular up to the muscle fiber level: Many productive processes of the lower level *together implement* a productive process at the upper level, and many individuals at the lower level *together comprise* the individual at the upper level. Each lower level serves as the explanans for the process at the level above it. Crucially, these explanations involve working components (the contractile elements mentioned) that bear joint role-filling relations. This means that these components jointly constitute the upper-level process, aligning with Gillett's concept of many-to-one compositional relations.

Two critical observations emerge from this analysis:

Firstly, the diversity of component parts, functions, and complexity of mechanisms varies significantly across the different levels of organization on which we focus our analysis. Examining all eleven levels in this example reveals that phenomena at some levels are of an aggregate

character, while others exhibit complex features such as multiple feedback loops, varying degrees of interactions, and intricate organization. This variability indicates that a system's complexity cannot be adequately judged at a single level. Instead, decomposition into subsystems and analysis of the phenomena they instantiate reveals diverse complexity profiles at each level. Consequently, while some levels are relatively easily decomposable, others present significant challenges in their decomposition and analysis. This observation leads to an important conclusion: the characterization of a system as decomposable or non-decomposable does not necessarily apply uniformly across all levels. At some levels of organization, the system may be readily decomposable, while at others, it may be only approximately so, or its decomposition may require a significant degree of idealization. This nuanced understanding of decomposability across levels is crucial for developing comprehensive explanations of biological phenomena.

Secondly, regarding the explanatory power at different levels, we observe a shifting dynamic as we move through the hierarchy. At higher levels with simpler organization, there's often a sense that the explanation for muscle contraction force is incomplete, prompting us to seek answers at lower levels. This perception changes as we descend the hierarchy. At the bottom level, where processes appear more complex and intricate, we encounter mechanisms that resemble tiny machines. However, it would be misleading to assume that these molecular mechanisms alone can fully explain why and how a muscle contracts and pulls on tendons. While reference to these lowest levels is necessary for a complete explanation, molecular mechanisms are insufficient for comprehending the phenomenon in its entirety. Each level contributes unique and complementary insights, collectively bringing the explanation closer to the observed macro-level phenomenon. This underscores a fundamental principle of compositional explanations: they are inherently multilevel. The role of decomposition is not merely to reach an explanation at the lowest level, but rather to facilitate a comprehensive understanding through the entire process of decomposition, the explanations formulated at each level, and the intricate relations between levels. In essence, the explanation emerges from the synthesis of insights across all levels, highlighting the integrative nature of compositional explanations in biology.

3.4. SUMMARY AND CONCLUSIONS

In this chapter, I explored the evolution of explanatory models in biological sciences, focusing on the development and application of compositional explanations. I began by examining the

limitations of traditional nomological and statistical approaches in explaining complex biological phenomena, highlighting the need for a more nuanced and multi-level approach. As we saw in the previous chapter, these phenomena are examined as systems characterized by aggregate complexity, which “concerns how individual elements work in concert to create systems with complex behavior” (Manson, 2001, p. 405). A special explanatory approach is needed, and this is achieved by Gillett's (2016) framework on scientific composition, which provides a robust theoretical foundation for understanding how entities at different levels of biological organization relate to one another in explanatory contexts and how they collectively constitute specific higher-level phenomena.

The elbow movement case study I presented served as a practical demonstration of how compositional explanations can be applied to complex biological systems. Through this example, I illustrated the process of decomposing a system across multiple levels of organization, from the whole-body level down to molecular interactions. My analysis revealed the importance of selective focus and guided decomposition in constructing effective compositional explanations. As we navigated through the hierarchical levels, from the broad question of forearm flexion to the specific mechanisms of muscle contraction, we encountered significant bifurcations that shaped our explanatory path. By this analysis I demonstrated that a comprehensive explanation is not an exhaustive pyramid of all possible compositional relations, but rather a targeted exploration of the explanatory relevant factors across multiple levels of organization. Furthermore, two critical observations emerged: First, the complexity and decomposability of the system varied significantly across different levels, challenging the notion of uniform decomposability. Second, the explanatory power shifted as we moved through the hierarchy, with each level contributing unique and complementary insights. This underscores the inherently multilevel nature of compositional explanations in biology, where the explanation emerges from the synthesis of insights across all levels.

This chapter conceptually leads to the fourth chapter, where the new-mechanistic explanatory approach will be presented and developed through examples. As we will see in the next chapter, explanatory mechanisms can have an etiological or constitutive character. In my thesis, I focus on explanations of the second type. These are constitutive (or compositional) explanations in the sense that a phenomenon or behavior of a system is constituted (or composed) by the collective action of working components that are described and examined at a lower level.

CHAPTER FOUR - DECOMPOSING PHENOMENA INTO MECHANISMS IN THE NEW-MECHANISTIC APPROACH

4.0. OUTLINE

In this chapter, building upon the foundations laid in the previous chapters, I explore the new-mechanistic approach as a species of compositional explanations utilized to explain biological phenomena. I aim to demonstrate how it synthesizes and extends the concepts of levels and decomposition in complex biological systems discussed earlier. I begin by tracing its development and application, particularly in the context of neuroscience (4.1. *Tracing the Roots of New-Mechanism*). I first define and generally describe the idea of New-Mechanism (4.1.1.) to recount, afterward, its historical roots, examining how it evolved from the ideas about hierarchical organization and decomposition (explored in the first two chapters of the dissertation) with a particular focus on the work of key philosophers that shaped the overall field (4.1.2.). Next, I present Craver's account of constitutive mechanistic explanation in neuroscience and address the crucial distinction between etiological and constitutive explanations (4.1.3.). In the second part of the chapter, I investigate any ambiguities in the definition and detection of a mechanism in the new-mechanistic sense (4.2. *What is and what is not a mechanism?*) examining different interpretations of mechanisms (4.2.1.). I, then, explore its boundaries and challenges through a discussion concerning borderline cases and the difficulties posed by highly complex biological systems (4.2.2.). In the final part, I present a detailed case study on spatial memory which involves mice navigating through a Morris water maze to provide a practical demonstration of constitutive mechanistic decomposition into a hierarchy of levels (4.3. *Mechanistic Decomposition: A Case Study in Spatial Memory*).

4.1. TRACING THE ROOTS OF NEW-MECHANISM

For the last 25 years or so, there has been a resurgence of mechanistic conception in explaining phenomena. This trend was particularly evident in life sciences, where due to their unique nature, the application of nomological explanatory models encountered insurmountable difficulties. The focus was on decomposing complex hierarchical systems to explain their phenomena

(compositional explanation), emphasizing the underlying mechanisms responsible for the causal processes at each level (causal mechanical explanation). In the second chapter, we saw that the idea of decomposition in analyzing a phenomenon was first proposed by Simon (1962), who applied it to complex hierarchical biological and economic systems. It was later enriched by Kaufmann (1970) and subsequently by Wimsatt (1972). Bechtel and Richardson utilized this idea in 1993 in the analysis of biological systems for explanatory purposes (2010), which led to the beginning of the new-mechanistic movement. An article by Stuart Glennan (1996) used the idea of *mechanism* to explain causal relations, and Machamer, Darden, and Craver's (2000) suggestion on the New-Mechanism idea led to the beginning of the now-called *New-Mechanistic Philosophy*.

4.1.1. Conceptualizing New-Mechanism

Let us examine the recent historical origins of the concept of mechanism. We can trace it back to the seventeenth century when it was used as one of the fundamental propositions of science to demonstrate that “all natural phenomena are *produced* by the mechanical interactions of the parts of matter according to mechanical laws” (Psillos, 2011, p. 773). This idea was rooted in Descartes's belief that mechanics was the basic building block of the physical world. The evolution of the concept of mechanism reflects the changing understanding of primary causal forces in the world (Craver & Tabery, 2019). While the current mechanistic conception in biological sciences still requires causal explanation, our understanding of what constitutes a mechanism has significantly evolved from the older, more concrete, machine-like concept. As discussed extensively in previous sections, biological systems are hierarchical and complex, both in terms of entities and processes, and in their organization across multiple levels. The new-mechanistic explanation emphasizes these features, which is why the proposed explanatory model treats these systems more broadly rather than as simple machines. According to Psillos (2011, p. 772), the contemporary concept of mechanism is “virtually *any* relatively stable arrangement of entities such that, by engaging in certain interactions, a function is performed or an effect is brought about.” Psillos distinguishes between two concepts of mechanism: a narrow concept deriving from the older mechanistic conception of nature, the *mechanical mechanism*, which refers to “configurations of matter in motion subject to mechanical laws,” and the currently used concept of pseudo-mechanical mechanism, *quasi-mechanical mechanism*, which is “*any* arrangement of parts into wholes in such a way that the behaviour of the whole depends on the properties of the parts and their mutual interactions.” The systems and phenomena examined in new-mechanistic philosophy belong to this latter category.

The applicability of this model in various sciences is under consideration. Machamer, Darden, and Craver (2000), who are among its main inspirers, believe that it can be extended to the general philosophy of science. However, there are significant challenges in applying a mechanistic model to physics, a fundamental science, while there is growing interest in exploring mechanisms in social sciences and economics. Additionally, there is an important exchange of ideas between mechanistic philosophy and computer and engineering sciences (Glennan & Illari, 2017a). Ultimately, the vast majority of the literature, as well as the two most comprehensive accounts of its application by Craver (2007) and Bechtel (2008), focus on biological (more specifically, neurobiological) systems, where mechanistic philosophy has thus far flourished. The subject of this dissertation also encompasses neuroscientific explanations, and therefore, the mechanistic accounts of Craver (2007) and Bechtel (2008) will be more extensively explored in this study.

The most fundamental principle of all new-mechanistic accounts is decomposing a phenomenon into smaller, constituent phenomena or decomposing the system that implements it into components and functions, along with a record of their spatial and temporal organization, i.e., a mapping of the interactions between the components of the system. The aim and outcome of decomposition are to reveal the hidden synthesis of a system. Here lies a significant difference that distinguishes a machine from a mechanism as is the latter defined in the new-mechanistic explanatory framework: while in a machine the usual intention is constructive, namely to design and construct parts that together carry out a particular operation, in a biological mechanism the intention is deconstructive, namely, to reveal its hidden synthesis, consisting of parts and functions, that comprise a system where a particular phenomenon is instantiated. This deconstructive goal is accomplished through the methodology of decomposition.

When discussing the component parts and functions of biological systems, particularly neurobiological ones, we must recognize that we are dealing with a fundamentally different paradigm than that of mechanical systems. In a machine, such as a spring-driven clock or a car engine, the components are typically well-defined, with clear boundaries and easily recognizable, discrete functions. These parts are concrete, observable objects that can be easily distinguished without specialized experiments or sophisticated methods. However, identifying parts and their functions in biological organisms, especially in neuroscience, is far more complex. According to Craver (2007), the systems studied in neuroscience often consist of components that are not fully visible or distinguishable within their environment. These parts are widely distributed, dynamically interconnected, and contribute simultaneously to various functions and phenomena. For instance, a

single neuron may participate in multiple neural circuits and cognitive processes. Furthermore, biological components often have context-dependent functions that can adapt over time (plasticity), and their boundaries are less distinct. This fundamental difference in complexity necessitates a more nuanced, holistic approach when analyzing biological systems. Unlike a clock's mechanism, which can be understood by simple observation and disassembly, understanding biological systems requires sophisticated experimental techniques and consideration of dynamic interactions, emergent properties, and the system as a whole.

An example of the enormous complexity encountered when one tries to detect parts in a neurobiological mechanism can be found in the field of molecular memory mechanisms research (Sweatt, 2007). Initially, researchers identified the hippocampus as the primary brain structure involved in memory consolidation. Through numerous electrophysiological experiments, they recognized a crucial component of the mechanism: the basic excitatory circuit. This circuit consists of three synapses that transfer information from the entorhinal cortex to the dentate gyrus and subsequently to the CA3 (cortical) area. Various diagrams illustrate this circuit and acknowledge it as part of the mechanism. Its characterization as a specific part (or module) is due to the robust stability of the connections detected between the involved neurons. However, it is also emphasized, and here lies the point of the example, that identifying this trisynaptic circuit is an oversimplification: other recurrent and collateral synapses between these cells contribute to and modify the mechanism. Even more so, we should consider that recognizing this circuit as part of the mechanism is currently a theoretical postulation of the prevailing theory of how memory works (Sweatt, 2007). This case is commonplace in neurosciences and differs significantly from how someone identifies the parts of a machine where parts are not indirect theoretical postulations waiting for experimental verification.

Craver (2007) notes that, for a neurobiological structure to be recognized as part (or entity) of a neurobiological mechanism, it must exhibit the following characteristics: "Real parts have a stable cluster of properties, they are robust, they can be used for intervention, and they are physiologically plausible in a given pragmatic context" (p. 131). The research methodology aims to determine whether the above applies to a possible entity of the mechanism. Consequently, the challenge of a mechanistic decomposition is to adequately exploit the experimental methodologies used in several biological and neuroscientific fields to reveal all the factors mentioned above.

In the new-mechanistic framework, a widely accepted definition of a mechanism is: “A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon (Glennan & Illari, 2017a, p. 2). While various other definitions have been proposed, decomposition remains a prerequisite for any mechanistic explanation. Despite differences in terminology, there is consensus that a mechanism for a phenomenon comprises entities and activities (Craver, 2007; Machamer, Darden, & Craver, 2000; Illari & Williamson, 2012) parts, functions, and interactions (Bechtel & Richardson, 2010), parts and operations (Bechtel, 2008) parts and interactions (Glennan, 1996), or entities (or parts), activities and interactions (Glennan & Illari, 2017b). These varied expressions all suggest that searching for a mechanism fundamentally involves identifying its component parts and functions, as well as their organization. In essence, any discussion of mechanisms necessitates consideration of decomposition.

It is worth noting that the rigor with which decomposition should be interpreted and applied in mechanistic explanations remains debatable. Questions arise: Must a system be completely decomposable to be mechanistically explained, or is the mechanistic ideal compatible with near-decomposable systems? Should all considered parts be concrete and material, or is the concept of mechanism applicable to more abstract systems? When addressing these questions, it is crucial to remember that this type of explanation is intended for biological systems, which are inherently complex. Attempts to decompose these systems according to specific criteria invariably involve approximations and idealizations in constructing explanatory models. Indeed, employing a degree of idealization is a common strategy for modeling and studying near-decomposable systems. In the following sections, I will examine the evolution of new-mechanism through its leading proponents. Subsequently, I will explore how broadly or narrowly a new-mechanism could be defined to encompass phenomena and systems in the special sciences. The final section will present the methodology of mechanistically decomposing a phenomenon in a neurobiological system, with reference to Craver's norms of mechanistic explanation.

4.1.2. The Evolution of New-Mechanistic Thought

While the roots of the New-mechanism idea are, as we have seen, considerably older, the initial prominent proposition for the present application of mechanistic explanation in biological sciences and psychology emerged from Bechtel and Richardson (1993). They advanced the notion that scientific discovery is facilitated by the strategies of decomposition and localization. They combined

Simon's work on the decomposition of hierarchical systems (Simon, 1962) and Wimsatt's study of complex biological systems (Wimsatt, 1972) with the scientific methodology in biology and psychology. The idea is that scientists harness complexity and construct mechanistic models for explanatory purposes by identifying causal components that explain how various systems produce specific phenomena. Their approach draws on scientific practice, which, as a human enterprise, is based on human reasoning. A consequence of the latter is that every scientific model is an approximation, or more accurately, an idealization.

The two elements of constructing an explanatory model, namely decomposition and localization, comprise a strategy that proceeds by analyzing a system into functional components with specific functions. According to Bechtel and Richardson (1993), "one of the strengths of decomposition and localization as a scientific strategy is that it facilitates an increasingly realistic representation of the explanatory domain, even when the initial representation is seriously distorted" (p. 8). This reflects the methodology of scientific discovery: the construction of evidence-based initial hypotheses that are gradually enriched and adjusted to constitute even more complete models of reality.

They introduced the term *mechanistic explanation* for understanding the behavior of complex biological and psychological systems, which they define as follows: "... these explanations, which we refer to as *mechanistic explanations*, propose to account for the behavior of a system in terms of the functions performed by its parts and the interactions between these parts" (Bechtel & Richardson, 2010, p. 17). Through this definition they provide a first identification of the elements that constitute a mechanism, namely *functions*, *parts*, and *interactions*; however, their inspiration for this explanatory model, at least in their initial approach to the subject, is not a biological complex system but the structure and function of manufactured machines. To be noted, their notion for such machines should not be interpreted based on the model of a typical machine like the mechanism of a clock; rather, their idea draws from a constantly changing machine's model that depends on the socio-historical context. Their definition of such a machine largely resembles the subsequent definitions formulated for the New-Mechanism; they describe it as a "composite of interrelated parts, each performing its own functions, that are combined in such a way that each contributes to producing a behavior of the system" (Bechtel & Richardson, 1993, p. 17). What they have in mind is complex machines with complex organizational principles such as multiple interactions with cascading effects, functional interdependence, feedback loops, and the like. These organizational structures are exemplified in their most extreme form within integrated systems.

In Bechtel and Richardson's (1993) initial formulation of the new-mechanistic account, the strategy of constructing a mechanistic explanation stands on two pillars: a bottom-up analytic approach, which detects the system's composition by isolating components physically within the system and then determining their role, and a synthetic top-down approach, which postulates different component operations for various subtasks within the system, collectively constituting the system's behavior. A prior hypothesis for the system's organization and operation is complemented with inhibitory and excitatory experimental methodologies that identify components and functions and detect behavioral effects of the components on the whole system, thus forming the necessary empirical constraints for the model.

At this point, the role of decomposition and localization enters the scene. *Decomposition*, to make the system more manageable, proceeds by breaking down the explanatory target, e.g., a function, in a set of subordinate functions, and *localization* ascribes the different subfunctions to the behavior or capacities of specific components. Several assumptions underlie these two heuristics of decomposition and localization. Among others, an important one is that the subfunctions are minimally interactive, which is an idealization. This assumption is connected with a second one about the system's decomposability. The implicit idea here is that the system is modular in character, i.e., "the components operate primarily according to their own intrinsically determined principles" (Bechtel & Richardson, 1993, p. 25). By this they mean, there is an operational closure among the components. They receive inputs and produce outputs received by other components of the same module and the module produces a specific output that is not strongly influenced by interactions with components residing in other modules or subsystems, or more generally, the module's function is not influenced by the system's function as a whole. When these two assumptions (or conditions), namely minimal interactivity and modularity, do not apply and there are many and various interactions among components, the system is characterized as integrated and minimally decomposable. Notably, these two heuristics can be employed only if a system is at least nearly (if not completely) decomposable, thus enabling a mechanistic explanation.

In Bechtel and Richardson's initial mechanistic account, interactions played a crucial role, often limiting the possibility of mechanistic explanations by increasing system complexity and integration. Within the systems tradition, interactions were previously used to define the boundaries of mechanisms and system decompositions based on the varying intensities of interactions within the system (Simon, 1962). As the concept of mechanism evolved, the notion of interactions became somewhat deflated, aligning with activities (or functions, or operations), since

most operations involve one thing acting upon another³¹. Practically, this means that it is an interaction. This shift is evident in Glennan's (1996) development of the mechanistic idea, where *interactions* are portrayed as carriers of causality between entities. In Glennan's account, and even more so in subsequent contributions to the new-mechanistic movement, the concept of interaction became dissociated from its role in increasing organizational complexity and determining decompositions. Instead, the focus shifted to its causal role. Indeed, while Simon (1962) initially proposed decomposing systems by comparing interaction intensities among different groups of component parts to separate subsystems, later approaches favored decompositions based on identifying component parts and functions, with interactions taking a secondary role. In contemporary mechanistic explanations, the role of interactions primarily pertains to organizing component parts, illustrating how the parts resulting from decomposition are structured to produce the phenomenon under study.

Glennan's first account that contributed to the development of the new-mechanism idea was not intended to show how explanations are achieved. The aim was to give a theory of what causality means and to reveal what is hidden behind the Humean "secret connection," i.e., Hume's regularity theory, which defined causation as the space and time contiguity of a cause and an effect, and the regular conjunction between those two. Through his mechanical theory of causation, he suggested that "two events are causally connected when and only when there is a mechanism connecting them" (Glennan, 1996, p. 64). In any other case, they are connected in virtue of a fundamental law. Glennan's definition of mechanism applies to complex systems that resemble machines, but it's clear he's not limiting his focus to the inner workings of literal machines. Instead, he's referring to complex systems with multiple underlying mechanisms driving various behaviors. This broad interpretation encompasses a wide range of phenomena, from human perceptual systems to market dynamics. His definition captures this usage: "A mechanism underlying a behavior is a complex system which produces that behavior by the interaction of a number of parts according to direct causal laws" (p. 52). Glennan keeps the terms of parts and interactions following Bechtel and Richardson's terminology.

³¹ Keiser (2017) notes that activities in most cases involve interactions. When they involve two entities, such as the sliding of the actin and myosin filaments past each other, they are binary. However, there are activities that involve only one entity, the so-called unary activities, such as the changing of the spatial formation of a protein. In that case, a unary activity (or operation, or function) doesn't entail an interaction.

Glennan's definition of a mechanism can be distilled into five key elements: a *complex system*, a *behavior*, multiple *parts*, various *interactions*, and *direct causal laws*. Drawing on Kaufman's (1970) ideas about decomposing complex systems, Glennan points out that such a system can perform multiple functions simultaneously, allowing researchers to investigate several mechanisms underlying different behaviors. He emphasizes the importance of clearly defining the target behavior before proceeding with decomposition. Significantly, the system's boundaries and its breakdown into parts and interactions are determined by the behavior being explained. However, Glennan stresses that these decompositions are not mere artifacts of description; rather, they reflect the genuine complexity and multifunctionality of the system.

Another point that Glennan makes clear is what kind of entities count as parts of a mechanism. He notes:

The parts of mechanisms must have a kind of robustness and reality apart from their place within that mechanism. It should in principle be possible to take the part out of the mechanism and consider its properties in another context. Care must be taken so that parts are neither merely properties of the system as a whole nor artifacts of the descriptive vocabulary. I shall summarize these restrictions by saying that parts must be objects.
(Glennan, 1996, p. 53)

Lastly, Glennan places particular emphasis on the concept of interactions that are governed by laws, specifically generalizations, which support counterfactuals. These should be direct causal laws. This requirement makes two points. Firstly, there is an actual interaction between parts only when laws are causal and not lawful generalizations explainable by common causes. Secondly, there must be direct interactions between any two parts without any other mediation (of parts or interactions). As he states, "a mechanism is a collection of parts, in which the behavior of the aggregate stems from a series of local interactions between parts" (Glennan, 1996, p. 56), and with "local," he means that there are no intervening parts. With his account and the mechanistic interpretation of the concept of causality, it was established that mechanical explanations are causal explanations.

The groundwork laid by Bechtel & Richardson and Glennan paved the way for Machamer, Darden, and Craver's (MDC) seminal contribution to mechanistic explanations. Their work established the foundation for what is now understood as *new-mechanistic explanation* (Machamer, Darden, & Craver, 2000). This study introduced notable changes in the terminology used to describe various aspects of mechanisms. The concept of interaction and its role in identifying causal relationships

was incorporated into the newly introduced idea of *activities* as change producers, while parts were reframed as *entities*³². Given the significance of their work in initiating and shaping the new-mechanistic movement, it is crucial to examine their definition of a mechanism, its constituent elements, and the concepts that contribute to its description.

MDC define mechanisms as “entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions” (Machamer, Darden, & Craver, 2000, p. 3). While this definition retains the core ideas and primary components of earlier conceptions—namely parts and functions/interactions—it expresses and describes them differently, introducing several new aspects and providing more detailed definitions. The key elements in this definition include *entities*, *activities*, *organization*, *productivity*, *regularity*, and *set-up* and *termination conditions*. Entities and activities, roughly corresponding to parts and functions or interactions in previous accounts, are the two fundamental ingredients of the new-mechanistic definition.

This account emphasizes certain key features of these elements. Entities possess specific spatiotemporal locations and are always individuated in terms of their properties, which enable their participation in specific activities. Entities and activities are complementary and interdependent: entities perform activities, while an entity's role as a component in a particular mechanism is defined by its involvement in specific activities. A subset of an entity's properties determines the activities in which it can engage. Like entities, activities have spatiotemporal locations but are also characterized by their rate, duration, and the types of entities and properties they involve. Activities represent what entities do; an entity engaged in an activity becomes a type of cause. Thus, a cause is an entity coupled with a specific activity, and activities are the producers of change. Each entity involved in a particular activity plays a functional role within the mechanism and comprises the context for the individuation of both entities and activities. Importantly, entities and activities can exist at any level of an examined mechanism, as they may form nested hierarchies with part-whole relations. Moreover, the MDC definition explicitly introduces the concept of organization. A mechanism's organization determines how its entities and activities produce the

³² For Glennan (2009), both his notion of interactions as causation bearers and MDC's notion of activities as producers of change retain an equal role of productivity, i.e., both make a general statement about causes: namely, that cause and effect are connected by intervening mechanisms. However, MDC place great emphasis on demarcating the new terminology they introduce as it contains subtle meanings that are very important for a thorough understanding of the new-mechanism.

phenomenon in question. Organization has both spatial and temporal aspects: for entities, these include location, structure, and orientation; for activities, they encompass order, rate, and duration. Organization also pertains to the stages of a mechanism, involving various organizational schemata and combinations such as lines, forks, joins, or cycles.

While the principle of organization was implicitly present in previous mechanistic accounts, MDC introduced several entirely new concepts in their work on biological mechanisms. One such concept is *productivity*. They assert that “an entity acts as a cause when it engages in a productive activity” (Machamer, Darden, & Craver, 2000, p. 6), linking productivity to causation and reflecting the capacity of entities and activities to transition a system from initial to final conditions. Specifically, activities involve either state transitions or the creation of new products. Another newly introduced feature is *regularity*: mechanisms typically function in the same way under identical conditions, implying that any state change is non-accidental and supports counterfactuals. This consistency is crucial for understanding and predicting mechanistic behavior. MDC also make a significant departure from previous models by eschewing the concepts of inputs and outputs. Instead, they introduce the notions of *set-up* and *termination conditions*. Set-up conditions encompass structural and spatial aspects, including relevant entities with their properties and various enabling conditions. Termination conditions, on the other hand, correspond to idealized states or parameters that describe a privileged endpoint, such as an equilibrium or an end product (which can be an entity or state). The processes within a mechanism are described through the intermediate stages of entities and their activities.

From this account onwards, the new-mechanistic movement experienced a unique flourishing. The terminology of activities, entities, and organization have largely dominated the subsequent definitions and descriptions of mechanisms. Carl Craver (2007) and William Bechtel (2008) are now widely regarded as the two main pillars in the philosophical literature related to the mechanistic explanation in cognitive science and neuroscience. Craver (2007), in his book on mechanistic explanation in neuroscience, gives the following definition of mechanism: “mechanisms are entities and activities organized such that they exhibit the explanandum phenomenon” (Craver, 2007, p. 6). Bechtel (2008) in his account, diverges slightly in terminology but maintains conceptual alignment. Rather than using “entities” and “activities,” he employs the terms “(component) parts” and “operations.” These concepts, within his conception of mechanism, correspond closely to their counterparts in Craver's definition: “parts” aligning with “entities,” and “operations” with “activities.” Specifically, for Bechtel, a mechanism consists of parts that are the structural

components, and operations that are processes or changes involving parts. He utilizes Bechtel and Abrahamsen's earlier definition: "A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel & Abrahamsen, 2005, p. 423).

Bechtel, in "Mental mechanisms: Philosophical perspectives on cognitive neuroscience" (2008) continued and enriched Bechtel and Richardson's (1993) original account with new elements. Two essential ideas stand out in his account: the application of mechanistic explanation to mental phenomena and his attempt to integrate these phenomena's dynamic and active character into his mechanistic explanatory model. Mechanisms are identified based on the phenomena for which they are responsible; namely, they are *mechanisms for certain phenomena*, and in a physical system, many mechanisms and phenomena may be identified. A phenomenon is a singular or repeatable occurrence, and it determines which are the parts, i.e., the structural components of a mechanism, and the operations, i.e., the processes or changes that involve parts. As is also evident from the treatment of entities and activities in MDC (2000) and related mechanistic accounts, parts and operations are complementarily defined and interdependent: a part is performing an operation, and an operation is performed on a part. Mechanistic decomposition is the procedure of determining which are the working parts and the operations of the mechanism, and in terms of parts, it is called *structural decomposition*, while for operations, it is called *functional decomposition*. Decomposition is completed with the localization procedure, meaning that every operation is related to a particular component part. One last feature that stands out in Bechtel's account is the so-called *phenomenal decomposition*, which is an attempt to demarcate and identify a phenomenon and leads to finding different types of phenomena that may compose it.

This brief review of the main contributors of the contemporary new-mechanistic account concludes with Bechtel (2008). In the following section, we will delve into Craver's new-mechanistic account in the context of the distinction between etiological and constitutive explanations. The latter type of mechanistic explanation is the focus of the dissertation. With this review, I attempted to outline the general picture of the new-mechanistic explanation. Although the definitions differ in their formulation and terminology, the idea is the same and is clearly articulated in the minimal definition proposed by Glennan and Illari (2017a, p. 2) according to which, "a mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon."

4.1.3. Craver's Account of Constitutive Mechanistic Explanation in Neuroscience

Craver's new-mechanistic account, which is the primary focus here, pertains to explanations in neuroscience: "Starting with neuroscience, as opposed to physics or chemistry, three main features of explanation demand attention: (i) explanations describe mechanisms; (ii) explanations span multiple levels; and (iii) explanations integrate findings from multiple fields" (2007, p. 2). Craver's general idea can be summarized in the statement that explanations in neuroscience are performed via the description of mechanisms that penetrate many levels and encompass many diverse scientific methodologies in a mosaic that unifies them under a complete explanatory schema (Craver, 2007). In his description of mechanisms, he emphasizes the divergence between biological mechanisms and machines or machine-like systems. This disparity arises from the fact that the component "parts of neural mechanisms are in many cases not so visible, not so readily distinguished from their surroundings; in some cases, they are widely distributed and dynamically connected, defying any attempts to localize functions to particular parts" (2007, p. 4). The mechanism's components are entities with particular properties such as location, size, structure, and orientation, enabling them to be involved in specific activities such as productive behaviors, causal interactions, omissions, and preventions. The particular spatiotemporal, causal and hierarchical organization of entities and activities is responsible for the specific phenomena presented by the system.

It's essential to highlight that in "Explaining the brain: Mechanisms and the mosaic unity of neuroscience" (2007), Craver provides a comprehensive description of how the decomposition of an explanandum phenomenon into the underlying mechanism that explains it is carried out. He also extensively examines what he means by "levels of mechanisms." This approach offers a mechanistic decomposition methodology, the application of which we'll explore in the final section of this chapter, using a spatial memory case study.

Craver carefully defines the significance of *levels of mechanism*, a concept that plays a pivotal role in his framework for mechanistic explanations specifically focused on how explanations are conducted in neuroscience. He highlights that, "I argue that 'levels of mechanisms' captures the central explanatory sense in which explanations in neuroscience (and elsewhere in the special sciences) span multiple levels. The multilevel structure of neuroscientific explanations is a consequence of the mechanistic structure of neuroscientific explanations" (Craver, 2007, p. 163).

Understanding his thorough treatment of levels of mechanism is crucial, as it forms the cornerstone for the constitutive nature of mechanistic explanations he presents for neuroscientific phenomena.

Craver's perception of mechanistic levels is very specific and specialized for each individual mechanism. He contrasts his level notion with Wimsatt's level prototype (1976; 1994). Wimsatt identifies different levels based on the size of objects that reside in a level, the compositional relations between them, laws and forces that govern them, the existence of regular and predictable relations in the same level, the theories developed in each level, and lastly, the techniques and disciplinary perspectives of the experimental methodologies in each level. Craver stresses that this categorization is quite general and, in many cases vague, since levels of a mechanism, according to his interpretation, are local in scope, serving the purpose of explaining specific phenomena. In any case, the point at which anyone who embraces the principles of systems tradition in defining levels would agree is that they regard part-whole relations and are not realization levels, as is the case, for example, with the three levels of analysis (computational, algorithmic and hardware/implementation) proposed by Marr (1976).

To define mechanistic levels, Craver (2007) poses three questions that could be a helpful heuristic for their identification: 1) what are the relata, 2) what is the interlevel relation, and 3) by virtue of what are any two items at the same level? Based on these three guiding questions, he proposes the following: They are levels of nature and not levels of science. Due to the interlevel relation of the relata, which is a part-whole relation, they are levels of composition. Lastly, although size and spatial overlap serve as a discriminating factor between entities of different levels, this is not the decisive criterion. What matters is the fact that the levels relate together acting entities that are working components in behaving mechanisms. Thus, he defines the interlevel relationship as follows: "X's φ -ing is at a lower mechanistic level than S's ψ -ing if and only if X's φ -ing is a component in the mechanism for S's ψ -ing. Lower-level components are *organized together* to form higher-level components" (Craver, 2007, p. 189). Notably, the notion of a mechanistic level is exclusively local, that is, it makes sense only within a specific mechanism. Namely, "X and S are at the same level of mechanisms only if X and S are components in the same mechanism, X's φ -ing is not a component in S's ψ -ing, and S's ψ -ing is not a component in X's φ -ing" (p. 192).

For Craver (2007), the mechanisms that explain neurobiological phenomena possess two important characteristics: Firstly, *causal relevance*. More specifically, the entities involved in the explanation of the phenomenon should be causally relevant to it. By building upon Salmon's (1984)

ideas, he argues that (good) explanations in neuroscience show how phenomena are situated within the causal structure of the world. In other words, an explanation should involve factors that play a specific role in causing the explanandum phenomenon. These factors are the components of a mechanism, i.e., entities and activities which are causally organized within the mechanism. Activities, as the bearers of causation “are productive”, “are not mere correlations”, “are not mere temporal sequences” and “they can potentially be exploited for the purposes of manipulation and control” (Craver, 2007, p. 6). Causal relevance is the primary concern in an etiological explanation as in the following: the release of neurotransmitters within a cell is triggered by a cascade of intracellular reactions that begin with the opening of calcium channels and end up in the formation of a pore between a vesicle containing the transmitter and the cell's membrane. This is an etiological explanation, meaning an explanation where an effect is explained through its antecedent causes.

The second characteristic that underscores Craver's significant contribution to the new-mechanistic movement is *constitutive relevance*. Entities participating in the mechanism underlying a given phenomenon should maintain a part-whole relationship with that phenomenon, namely, they should constitute it. This characteristic is related to Craver's emphasis on the multilevel character of explanations in neuroscience. The general idea of constitutive relevance can be summarized as follows: a phenomenon occurs within a system composed of numerous component parts that interact in various ways. A mechanism is utilized to explain a specific phenomenon and only that, given that the same system could also give rise to other phenomena explained by different mechanisms. Consequently, not all component parts and interactions within the system pertain to the specific mechanism. As Craver states, “Good mechanistic explanatory texts describe all of the relevant components and their interactions, and they include none of the irrelevant components and interactions” (Craver, 2007, p. 140). Hence, every experimental strategy aimed at identifying components, i.e., entities and activities of a mechanism, must exclude all irrelevant to the mechanism parts of the system. For Craver, the criterion of constitutive relevance for a part lies in mutual manipulability, as defined: “A part is a component in a mechanism if one can change the behavior of the mechanism as a whole by intervening to change the component, and one can change the behavior of the component by intervening to change the behavior of the mechanism as a whole” (2007, p. 141). From this perspective, Craver introduces yet another evaluative criterion for assessing the adequacy of an explanation underlying a phenomenon.

The constitutive dimension of a mechanism is not a universal aspect present in all mechanistic explanations. Some mechanisms describe the production of a phenomenon by explaining what precedes it without presupposing its analysis into entities and activities at a lower level of composition. In other words, not all explanations have a compositional character in the sense discussed in the previous chapter; they are not “explanations of higher-level entities built around lower-level entities taken to compose them” (Gillett, 2016, p. 2). The definition of mechanism that MDC (2000) give in what is widely considered the seminal paper of the new mechanistic movement is clear in this regard: A mechanism consists of “entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions” (Machamer, Darden, & Craver, 2000, p. 3), and it does not explicitly state whether these entities and activities constitute the phenomenon at a lower level.

The latter brings forth an important distinction that concerns mechanistic explanations: the distinction between *constitutive* and *etiological* mechanistic explanations. It was initially coined by Salmon (1984) to distinguish between two aspects of causal-mechanical explanation. The etiological aspect looks back and reveals the antecedent causes of an event, while the constitutive aspect looks within the event to identify its components and their organization. Craver points out this distinction and clarifies that the norms for mechanistic explanation he proposes concern constitutive explanations:

Explanations in neuroscience describe mechanisms. Some mechanistic explanations are etiological; they explain an event by describing its antecedent causes. Dehydration is part of the etiological explanation of thirst. Prion proteins are part of the etiological explanation of Creutzfeldt-Jacob disease. Excessive repetition of the CAG nucleotide pattern on the fourth chromosome is part of the etiological explanation for Huntington’s disease. Other mechanistic explanations are constitutive or componential; they explain a phenomenon by describing its underlying mechanism. The NMDA receptor is part of the constitutive explanation of LTP. The hippocampus is part of the constitutive explanation for spatial memory. Ions are part of the constitutive explanation for the action potential. In this chapter, I develop a normatively adequate account of constitutive explanation. (Craver, 2007, p. 107)

An example provided by Craver, Glennan, and Povich (2021) further illustrates this distinction. In the example, the locomotion of a worm is explained in terms of biological mechanisms involving

muscles, neurons, neuronal networks, receptors, and ions. The touch receptors on the worm's head transform a mechanical stimulus into an electrical signal, which is then transmitted from sensory neurons to motor neurons and muscles through neurotransmitters. According to Salmon's distinction, the etiological aspect of the explanation focuses on understanding the worm's locomotion by tracing back and revealing the antecedent causes that led to the event. On the other hand, the constitutive aspect looks within the worm to identify its internal components and their organization, explaining how they interact to bring about the phenomenon. Notably, in etiological explanations, we observe an emphasis on temporal relations: on the sequence of events that precede the occurrence of the phenomenon, while correspondingly, in constitutive explanations, the emphasis is on spatial containment: on the entities and processes within the system that instantiate the phenomenon.

Kaiser and Krickel (2017) explore the metaphysics of constitutive mechanistic phenomena and note that constitutive and etiological explanations differ in their grounding relations. A constitutive explanation is grounded in a non-causal relation; constitutive mechanisms and the phenomenon are related by mechanistic constitution, which is not causal; it concerns parthood and gives rise to several levels of organization. On the contrary, etiological explanations invoke causal relations between preceding causes and their effects; the mechanism causes the phenomenon entirely (Kaiser & Krickel, 2017). Importantly, the authors emphasize that a mechanism cannot be simultaneously etiological and constitutive; even more so, the same phenomenon cannot be explained by both types of mechanistic explanation. They illustrate it by referring to the example of the protein synthesis explanation. An etiological explanation for protein synthesis would describe the series of causes that lead to the synthesis of a protein, while a constitutive explanation would describe the components of a cell and how they act and interact to enable protein synthesis. However, while the two explanations might seem to be explaining the same phenomenon, they are actually addressing two different aspects of it: the etiological explains the end-product of protein synthesis, which is the presence of a protein, while the constitutive seeks to understand what happens at every step of protein synthesis, namely, the process of protein synthesis.

4.2. WHAT IS AND WHAT IS NOT A MECHANISM?

Thus far, we have examined how the most prominent accounts have attempted to define and describe the idea of New-Mechanism. A common factor in all of them is that they refer to

mechanisms “for a phenomenon” and require the decomposition of the phenomenon or system that substantiates it into a number of temporally, spatially, and causally organized entities and activities or interactions. In what follows, I will consider two issues that will complement the definition of a mechanism within the new-mechanistic framework. The first concerns the varied meanings of the term “mechanism” in science and philosophy. It is crucial to identify these different interpretations to distinguish between new-mechanism and other conceptualizations of mechanisms. The second issue pertains to the scope of mechanisms – specifically, which phenomena and systems can be effectively studied and explained using mechanistic approaches. The exploration in these two sections is closely tied to the widely accepted definition of mechanisms which states: “A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon” (Glennan & Illari, 2017a, p. 2).

4.2.1. Distinguishing Mechanisms: Various Interpretations

Firstly, let us delve into the various ways the concept of a mechanism is employed in the philosophy of science. Andersen (2014a; 2014b) identifies five contemporary usages of the notion. The first usage pertains to the concept of the new-mechanistic idea as it is integrated into scientific practice. This first understanding of mechanisms is then juxtaposed with four alternative interpretations that do not align precisely with the essence of a new-mechanistic explanation. Identifying and demarcating them helps clarify what we mean and do not mean when examining a phenomenon under the new-mechanistic framework. Simultaneously, recognizing these different facets of mechanistic interpretations and understanding their frequent association with disputes in the philosophy of science underscores the importance of precisely defining the term “mechanism” within the context of the new-mechanistic philosophy framework. Ultimately, as Andersen points out, it appears that these various senses share no substantive core meaning.

According to Andersen (2014a), the first sense of mechanism (mechanism₁) is expressed mainly through the new-mechanistic movement in the philosophy of biology. With mechanism₁ she refers to *mechanisms as integral to scientific practice*. A key element is the presence of a hierarchical structure with nested levels consisting of organized entities, activities, and interactions, and this structure is reflected in the scientific methodology and the explanations provided. As Andersen particularly emphasizes, this sense of mechanism simultaneously provides two types of explanations: a horizontal one, related to interlevel causal relationships where the explanandum is

the final stage in an often complex causal chain and a vertical where the mechanism's operation constitutes the explanandum at the higher level³³.

When discussing this type of mechanism, we imply certain claims about the ontology, methodology, and representation of phenomena that share a common hierarchical structure. It's important to clarify that any similarity between the ontology and the representation of a phenomenon doesn't automatically mean that the explanatory model, which serves as a means of explaining it, perfectly matches the phenomenon itself or possesses identical attributes. The phenomenon is inherently more complex, and while features such as its hierarchical structure might pertain to both the phenomenon and the representation, the explanatory model remains an idealized abstraction. It encapsulates its fundamental characteristics with a certain level of approximation.

Consider, for example, the model of muscular force production introduced in the previous chapter, which describes a mechanism at the subcellular level. In this model, myosin necks form cross-bridges with actin-binding sites, exerting a pulling force that propels them in parallel towards the sarcomere's central region. This configuration resembles a miniature machine. However, this machine-like image of entities and activities is, to some extent, an idealization, as the molecular processes at this level are far more probabilistic and less defined than the model suggests. The model aims to represent only a set of particular mechanism features, assuming that they play a leading role in its function. Nevertheless, it is not far from reality, especially regarding the hierarchical and nested structure of the involved mechanisms. The model effectively balances simplification with accuracy to provide a useful representation of the underlying processes.

Another characteristic of mechanism₁, as a property of the new mechanism, crucial for its explanatory capability, is regularity. As pointed out by Andersen: “This brings us to a key feature of mechanisms₁ by which they can be distinguished from other senses, especially from mechanisms₂. Mechanisms₁ are characterized in part by regularity. Mechanisms₁ are the kinds of things that do or straightforwardly could recur” (Andersen, 2014a, p. 279). This feature is what grants the new-mechanism its explanatory power, suggesting that mechanistic descriptions and explanations possess a generality that can be extended to other phenomena of a similar nature.

³³ Clearly, Andersen here refers to the distinction between etiological and constitutive mechanistic explanations that we explored in section 4.2.3.

Andersen introduces a second sense of mechanism, mechanism₂, which she defines as *mechanism as an ontology of the world*. In mechanism₂, we observe both similarities and differences compared to Mechanism₁. In particular, here, too, we find hierarchically structured layers of mechanisms comprised of smaller mechanisms with organized entities and activities, but this hierarchy picture makes an ontological claim about what exists in the world and provides a complete account of causation. Unlike the first sense of mechanism, which concerns types of phenomena, the second sense regards tokens. This brings forth the importance of regularity as an essential feature of an explanatory mechanistic description: as we have already seen in the MDC (2000) account of the New-mechanistic movement, regularity, as a newly introduced feature of a new-mechanism, provides a nomological validity to the explanation, since any occurring state change is non-accidental and supports counterfactuals. A consequence of this is that a new-mechanistic explanation can be applied to many tokens of the same type's phenomenon. But the second sense of mechanism does not need regularity to be explanatory useful because it is meant to explain only a token of a phenomenon.

Andersen (2014b) outlines three additional interpretations of mechanisms in the philosophy of science, which are semantically more distinct from the earlier ones primarily due to their flat nature. Unlike the previous interpretations, these, namely mechanism₃, mechanism₄, and mechanism₅, do not inherently assume a hierarchical structure.

The third interpretation of mechanism, mechanism₃, involves the utilization of the term “mechanism” to denote *single equations within a structural equation model*. In this sense, each equation serves as a mathematical representation of the causal relations among variables within a system. A defining attribute of this interpretation is independence; in essence, the system's representation necessitates a modular structure. This implies that the representation comprises several autonomous equations, each of which can be examined in isolation from the overall equation. Mechanism₃ is quite distant from the new-mechanism, (referred here, as mechanism₁). For instance, Hodgkin and Huxley's model (1952) consists of four individual equations that correspond to four individual currents (the capacitive current, the potassium current, the sodium, and the leakage current). All these constitute the total current crossing the membrane, the intensity of which determines the creation of action potential in neurons. Although the system's behavior is decomposed into sub-behaviors, this decomposition is quite different from mechanistic decomposition. In mechanism₁, there are part-whole relations between the lower and higher-level

components; a system is decomposed to reveal the lower-level mechanisms that explain the upper-level behavior. Hence, the similarity regarding decomposition is only superficial.

Next, Andersen introduces the fourth notion of mechanism, mechanism₄, which, like mechanism₃, is also linked to the concept of causation. She defines it as *mechanisms as processes and/or interactions in physical process accounts of causation*. Here, we observe a clear differentiation from mechanism₁. Specifically, mechanism₄, inspired by Salmon's Causal Mechanical account (1984), has a limited scope of application as it refers specifically to causal mechanisms that explain phenomena belonging solely to the physical and material domain. Consequently, explanations concerning social, historical, and similar phenomena are excluded. This stands in contrast to mechanism₁, which can be applied to these fields as well. More importantly, this sense of (causal) mechanisms can not accommodate omissions and preventions³⁴ that are very common in biological sciences, as Craver (2007) has extensively argued, and this is an essential methodological difference that distances mechanism₄ from mechanism₁.

Last but not least, the fifth sense of mechanism, mechanism₅, which should be distinguished from mechanism₁, regards *mechanisms as net constraints on the states open to a system*. This interpretation significantly diverges from all the preceding ones, primarily because it pertains to information-theoretic systems and constitutes a mathematical characterization in a purely abstract system. This mechanism is a pattern of constraints that determine the possible states in which an information system may find itself; in other words, it describes and predicts its repertoire. Mechanism₅ is not explanatory, so it is even more distant from all the previous senses.

So far, we have seen that when the term mechanism is used to characterize, describe, and explain a phenomenon, what exactly is meant can vary as its reference depends on entirely different conceptual frameworks within the scope of the philosophy of science. Hence, we see that the idea of the new-mechanism not only is distanced from the older use of the term concerning machines but there is also a conceptual confusion in the contemporary use of the term. Andersen (2014a; 2014b) adequately addressed this conceptual confusion, excluding uses of the term that do not really refer to what is meant within the new-mechanistic movement. So, when new-mechanists refer to a mechanism, they mean the explanatorily and methodologically useful exploration of a levels

³⁴ Omission is when the absence of a cause allows an effect and prevention is when a cause inhibits or precludes an effect.

hierarchy where the lower levels through their entities' operations can explain the phenomena observed at the higher, and this can be generalized to similar phenomena. The potential for mechanistic decomposition plays a dominant role in the construction of such an explanatory model. Ideally, in the case of constitutive mechanistic explanation, this would result in identifying entities and activities at a level lower than that of the phenomenon being explained. However, it is important to keep in mind that a mechanism constitutes an explanatory model and as such, involves a certain degree of approximation and abstraction from the phenomenon itself. Of course, tolerance for approximation is not unlimited and while some descriptions might appear to be mechanisms, they fail to meet the criteria for being labeled as such. On the other hand, certain phenomena may not be explainable as mechanisms for various reasons, the primary one of which, is the potential for decomposition into the constituent elements comprising a mechanism. This narrows the scope where mechanistic explanation can be applied.

4.2.2. Boundaries and Challenges of Mechanistic Explanation

What we consider to be a mechanism, within the framework of the new-mechanistic movement, depends firstly on how we interpret its definition and secondly on which behaviors and systems we assess to fall within its scope. We discussed the first part of this consideration (various interpretations of defining mechanisms) in the previous section. The second issue I want to examine in the present section regards which phenomena and systems could be studied and explained as mechanisms. One way to explore this is by examining borderline cases, meaning cases that could almost be characterized as mechanisms but are not. This approach allows us to investigate the scope and limitations of our definition's applicability. A second way relates to what is required for considering a system to possess spatially and temporally organized entities, and activities and will be examined thereafter.

In what follows, I will begin to explore this topic by examining six borderline cases presented by Illary and Williamson (2012). These cases closely resemble a mechanism but are not classified as such for several reasons that have to do with the essential features of a mechanism, as it is construed in the new-mechanistic account.

The first case regards scientific advances that may yield partial mechanism descriptions, but these descriptions fail to serve as models that explain the relevant phenomena properly. For instance, as seen in the case of discovering various memory forms, like long-term, short-term, working, episodic, semantic, implicit memory, and several forms of priming, while separate mechanisms

could produce these, we cannot yet even determine their number. Certainly, this categorization of memory types corresponds to a more accurate description of the phenomenon, but only uncovering underlying entities validates separate mechanisms, indicating that prior to this, our understanding is limited to partial descriptions. A second borderline case is the existence of entities without activities. This case could refer to a fixed machine, such as a system of pulleys and levers, but not to a mechanism that is, by definition, active. The opposite boundary case, which refers to activities without entities, corresponds to a functional explanation without realization. This means, according to Bechtel's (2010) account, it is a functional decomposition where localization of operations is missing, or as Craver (2007) puts it, is not a mechanism schema or model but just a mechanism sketch without any details for the material parts that realize the mechanism. Hence, as most new-mechanists insist, entities and activities are complementary, and if any concrete or abstract structure lacks one of the two, it is not a mechanism.

The fourth boundary case according to Illary and Williamson (2012) pertains to the element of organization and the scenario where it is either absent or so complicated that it cannot be traced and represented in a comprehensible manner. In other words, there is no way to describe entities in space and activities in time in a coherent way that makes sense. Nonetheless, even in such cases, ultimately, the generation of the phenomenon suggests the existence of some form of underlying organization. A fifth case is when a mathematical description effectively explains and predicts a phenomenon, yet this description lacks concrete entities, thus not constituting a causal mechanism. While it may correspond to a mechanism, this mathematical representation falls short of being one. The final, sixth boundary case, concerns the balance between model accuracy and idealization in describing a phenomenon. These attributes are interconnected, with the extent of idealization in a model influencing its accuracy. Even if an idealized model demonstrates predictive capabilities, deviations or significant disparities from the actual phenomenon prevent it from aligning with reality, thereby also being a marginal case of a mechanism.

In the scenarios described earlier, some key attributes of mechanisms were not fully realized. These attributes are essential for providing a comprehensive mechanistic explanation of a phenomenon within a system. Consequently, these cases fell short of meeting the criteria necessary to be classified as mechanisms. When evaluating these situations and attempting to establish the essential criteria for classifying a phenomenon as mechanistically describable and explainable, it is important to strike a balance between two extremes. On one hand, an overly broad definition would lack specificity and would too liberally categorize many borderline cases as mechanisms. On the

other hand, an excessively narrow definition would fail to encompass the wide variety of behaving systems that genuinely qualify as mechanisms.

Especially in the case of biological systems, this difficulty is particularly evident. If we focus our reasoning on biological systems, we will realize that we are dealing with living organisms, which are active mechanisms and not mere machines like clocks whose parts and functions would be more or less easily detectable. Occasionally, identifying specific anatomical entities and their activities is relatively straightforward, provided they possess distinct attributes that set them apart. Nevertheless, this is not always the case. Often, the situation becomes quite challenging when attempting to mechanistically analyze more intricate internal structures at various size scales. Serious difficulties arise when we try to break down and identify the entities, activities, and their organization within such systems. This is particularly true for systems responsible for functions that we do not fully comprehend yet. In these cases, the process of mechanistic analysis becomes significantly more complicated and less evident.

That is when the model should balance between approximation and precision, with the second requiring identification of those appropriate elements that will enable it to faithfully reproduce specific behaviors of the system under study and predict other behaviors that may arise. Even then, however, it should be kept in mind, that the identified component entities and activities resulting from the decomposition of the system and the specific phenomenon that we study, are mainly attributes of the explanatory model, not of the phenomenon and the system per se. Given all the above, the dilemma we face is deciding whether the behavior of a system can be modeled and explained as a mechanism, even with a certain degree of approximation, or if there are compelling reasons that make it impossible. In this line of thought, Glennan and Illary's (2017a) definition of a mechanism³⁵, has an adequately broad scope.

In further defining the varieties of a mechanism, Glennan and Illary (2017b) focus on several taxonomic characteristics. Following their suggestions and trying to deepen our understanding of mechanistic decomposition, we will focus on their remarks about entities, activities, and organization. Entities and activities are the end product of the decomposition procedure. Entities are the material working parts of the mechanism, and they can widely vary in terms of kind, shape,

³⁵ "A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon (Glennan & Illari, 2017a, p. 2)

size, and other features. Their definition is intrinsically linked to their functions, as an object only qualifies as an entity within a system when it participates in a specific activity or interaction. Conversely, an activity is always carried out by a particular entity³⁶. This relationship of mutual definition between entities and activities is not determinative; it just means that they constrain each other.

The authors argue that the characterization of something as an entity or activity is local, namely, in the context of specific phenomena. This locality has certain implications. In the first place, something is considered a working component provided that it somehow relates to the mechanism within which it is located, and this relation is determinative for identifying the working component. Given that it possibly contributes to another mechanism as well, it may be defined differently there. After all, entities are bearers of properties that allow them to engage in specific activities, and these properties define the activities they take part in. A relevant biological example of an entity that can be defined as part of two mechanisms simultaneously is the cell membrane: Due to its property of being composed of two layers of lipids, the cell membrane takes part in the mechanism that sets the cell's boundary and separates it from the extracellular space. It also possesses the property of having controlled permeability to certain types of ions, and therefore, it serves as an essential working component in the mechanism for the transmission of action potentials.

Another point of interest about entities that the authors consider has to do with their ontology. Are the discovered working components real entities of the world? The authors keep a cautious realistic view of this matter, but they note that what counts as an entity (or working part) is often context and model-dependent. They point out that the same mechanism can include entities of notably different sizes, suggesting that what counts as a working component depends on the scale and perspective of analysis and emphasize that there is no absolute acknowledging of a real entity or working component. Along the same line, others like Kaiser (2017) and Kaplan (2008) note that since scientists perform decomposition for explanatory purposes, the discovered components may be artifacts of the explanation and the discovery procedure. In this sense, there may be a discrepancy between the real components belonging to a system that implements a mechanism and

³⁶ To be noted, interactions are activities that involve more than one entity. According to Kaiser (2017), some activities involve only one entity, such as changing the proteins' spatial conformation. This case is an activity that is not an interaction. Therefore, it is reasonable to refer to them with separate terms, i.e., activities and interactions; nevertheless, the term activity could be used for both cases.

the components belonging to its possible model; this could result from a lack of understanding of the phenomenon's actual mechanism. However, since the new-mechanistic explanation does not make any explicit ontological statement, insofar as the model represents the phenomenon and the system relatively accurately, this discrepancy is expected and does not undermine the value of constructing mechanistic models for explanatory and predictive purposes. In any case, the upshot of all these is that when the subject matter is decomposition in complex biological systems, there is always some degree of approximation and idealization involved.

Another point to consider is the organization's role in characterizing a system as mechanistically explicable. Organization of a system's components is one of its essential features; if two systems have the same components but are differently organized, their behaviors and characters will be different. For instance, considering the organization pattern of muscle fibers in a muscle, namely the muscle architecture, we will discover important consequences regarding its ability to shorten and exert traction to a bone. A parallel muscle is typically composed of long muscle fibers that have a length equal to the length of the entire muscle and are arranged parallel to each other and parallel to the muscle's longitudinal axis. Due to this organization, the muscle can contract to a large extent and thus move the joint it attaches to over a wide range. On the other hand, a pennate muscle usually consists of a central long tendon and multiple shorter muscle fibers that attach to it, resembling the structure of a feather. Such muscles, because of their feather-like organization, can generate significant contracting power while sacrificing their ability to shorten. We see here how the organization of these entities, the muscle fibers, influences the behavior of the system as a whole. Each muscle fiber, as an individual unit, functions in a specific way, but their different organization in different muscle types results in the varied function of the entire system (Oatis, 2009). Here, the factor that contributes to the greatest extent in explaining the phenomenon described above is the organization of muscle fibers within the muscle.

Organization plays a crucial role in determining a system's decomposability too, and consequently, its amenability to mechanistic explanation. Systems with clearly defined organizational structures - such as distinct component clusters in specific spatial locations or well-delineated local and functional interactions - lend themselves to analysis and description through mechanistic models. However, as organizational complexity increases, incorporating features like positive and negative feedback loops, multiple recursive connections, and varied interactions between entities (which may simultaneously participate in diverse processes with different roles and intensities), the mechanistic description becomes more challenging. Nonetheless, this more complex picture aligns

closely with the nature of living and cognitive systems, as complexity is a hallmark of both intelligence and life itself.

In direct relation to this issue is the critique of the mechanistic philosophy posed by Silberstein and Chemero (2013) earlier and Silberstein (2021). This critique focuses on the use of localization and decomposition in the mechanistic explanation of complex biological systems. They argue that historically mechanistic explanation was defined exhaustively in terms of localization and decomposition, and till now, these are considered both necessary and sufficient conditions for this kind of explanation. However, causal and spatiotemporal organizational features of complex biological mechanisms in the real world such as global constraints and other kinds of context sensitivity violate localization and decomposition. They support their position with examples from various fields of systems biology, such as genetics, epigenomics, molecular biology, developmental biology, and synthetic biology. In these fields, properties like plasticity, robustness, autonomy, and universality possessed by biological systems challenge the notion of hierarchical, compositional, and constitutive explanations advocated by the new-mechanists.

Silberstein (2021) then considers as one possible solution for this problem the rejection of localization and decomposition as essential for a mechanistic explanation. In this case, he questions what remains as the essence of mechanistic explanation if localization and decomposition are abandoned, suggesting that the new-mechanistic philosophy might become too broad or trivial. Nevertheless, Silberstein and Chemero refer to a strict interpretation of decomposition, but as we have seen, many new-mechanists reject this historical machine-like conception. In any case, Silberstein and Chemero (2013) and Silberstein's (2021) concerns make the question of how broadly or narrowly we should interpret the notion of a mechanism in explaining biological phenomena more urgent. Decomposition is a hallmark of mechanistic explanation and its interpretation within the new-mechanistic framework should be carefully treated. Hence, we must interpret decomposition cautiously.

Keeping the remarks mentioned above in mind, we can now have a closer look at a case of mechanistic decomposition as it is performed, aiming to investigate, describe, and explain a phenomenon in the brain. In what follows, I will proceed to the methodology of the mechanistic decomposition of a neurobiological system that instantiates a mental phenomenon. For this purpose, I will mainly refer to the norms and diagrams used by Craver (2007), who has taken into account all the unique features of phenomena and systems belonging to the field of neuroscience.

4.3. MECHANISTIC DECOMPOSITION: A CASE STUDY IN SPATIAL MEMORY

In this section, exploiting the norms of constitutive mechanistic explanation in levels of mechanisms proposed by Craver³⁷ (2007), I will explore the methodology of mechanistic decomposition for explaining a phenomenon and a system utilizing the case of spatial memory, which is a widely investigated phenomenon whose study unfolds at many levels of organization.

The process of decomposing a phenomenon begins with a fundamental question: How is it produced? To answer this, an explanatory model is developed that outlines the various mechanisms involved at different levels of organization. For this to be accomplished, it is necessary to identify the entities and activities that contribute to the phenomenon being studied. A clear definition of the phenomenon is crucial to this process. It is also important to identify the appropriate hierarchical levels where entities and activities are located. These levels are not predetermined and are detected by the decompositional methodologies used. The identification of these levels is based on scientific assumptions and methodologies that form the theoretical foundation of each discipline. For instance, when studying a cognitive function in a particular region of the brain, the involved entities will be found at the level of neural networks or local structures. This local nature of mechanistic levels is reflected in Craver's definition of interlevel relationships: "X's φ -ing is at a lower mechanistic level than S's ψ -ing if and only if X's φ -ing is a component in the mechanism for S's ψ -ing. Lower-level components are *organized together* to form higher-level components" (Craver, 2007, p. 189).

Craver describes and thoroughly analyses the application of the mechanistic explanatory model to neuroscientific research. He notes that discovering the mechanism underlying a phenomenon requires decomposing it into entities and activities and capturing those elements' organization. As is shown in his schema (*Figure 9*), a phenomenon is a system *S*, which exhibits the behavior ψ -ing,

³⁷ The relevant literature will reveal several attempts to apply the mechanistic explanation in the mental realm. The two leading positions draw from Craver's methodology (Piccinini & Craver, 2011) or Bechtel's (2008). To be noted, Bechtel's (2008) ideas about structural and functional decomposition and localization are relatively close to Craver's account, although he uses different terminology. Craver's (2007) work on the norms of mechanistic explanation, his detailed explanation of mechanistic levels, and his comprehensive representation of his methodology through relevant schemata make his model particularly sophisticated and comprehensive. Therefore, I will adopt his framework for the methodology of decomposition.

e.g., it does ψ . The system is analyzed in component parts x_1, x_2, x_3 , e.ct., and each part is engaged in an activity $\varphi_1, \varphi_2, \varphi_3$, e.ct. In Craver's schema, the parts are connected with arrows indicating the causal interactions with each other. Each of these parts can be examined at a lower hierarchical level, where a similar type of analysis can be applied (*Figure 10*). This reflects the locally nested (*Figure 11*) hierarchies of subsystems that perform sub-functions and constitute part of the phenomenon's explanation, which, as he states, “oscillates up and down in a hierarchy of mechanisms” (Craver, 2007, p. 10).

Entities and activities alone cannot explain a phenomenon if not accompanied by an account of their organization. In line with Wimsatt's (1986) ideas about the role of organization in shaping a complex system's properties and behaviors, Craver (2007) stresses the importance of organizational principles, which he categorizes as active, spatial, and temporal. The first category, i.e., active organization, regards the cooperative and competitive interactions between the parts of a working mechanism, spatial organization includes sizes, shapes, structures, locations, orientations, directions, and connections of parts, and temporal organization is about order, rhythm, and duration of activities. This is the general idea of Craver's mechanistic decomposition into organized entities and activities that extend into a hierarchy of nested levels.

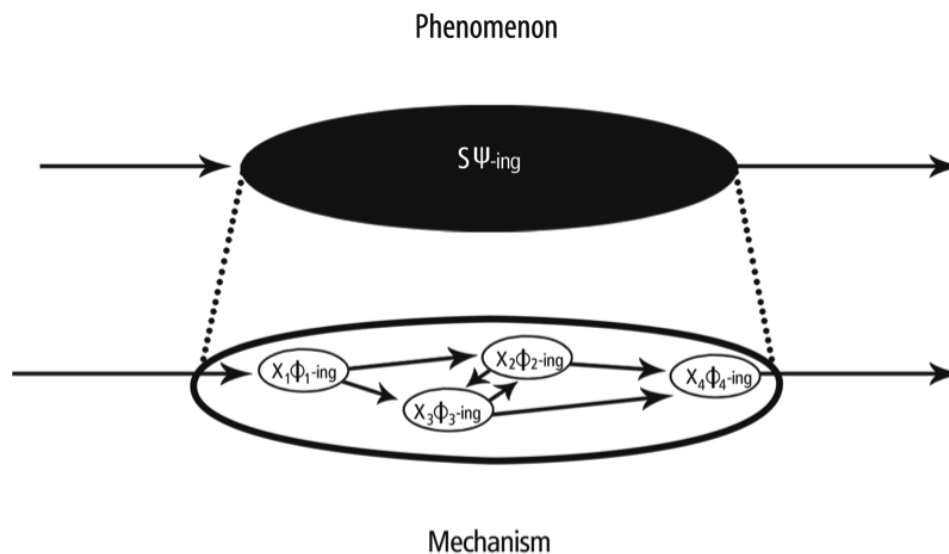


Figure 9: A phenomenon (top) and its mechanism (bottom). Reproduced from Craver (2007, p. 7)

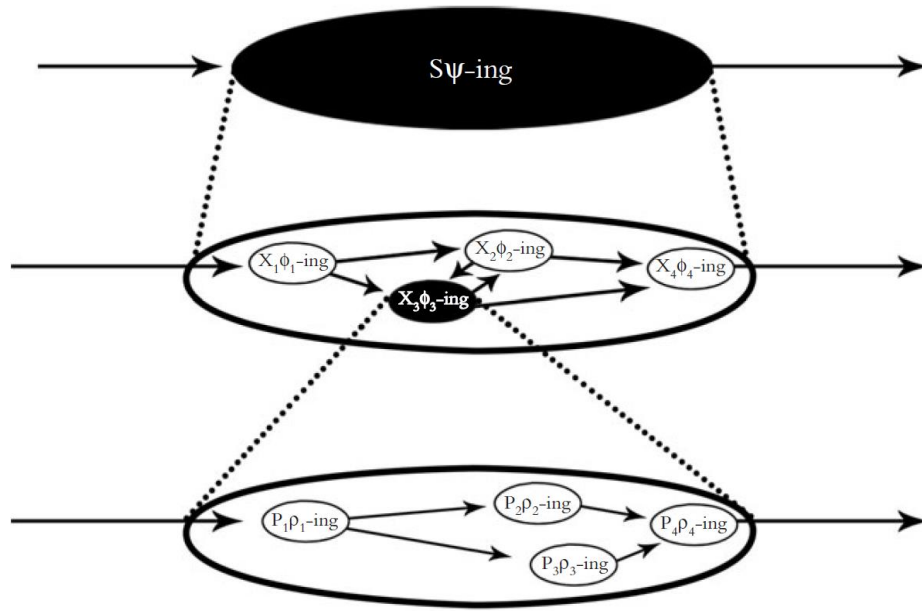


Figure 10: Three levels of mechanisms. Reproduced from Craver (2007, p. 189). Lower-level parts constitute higher-level parts through spatial, temporal and active organization into something larger than the sum of the parts.

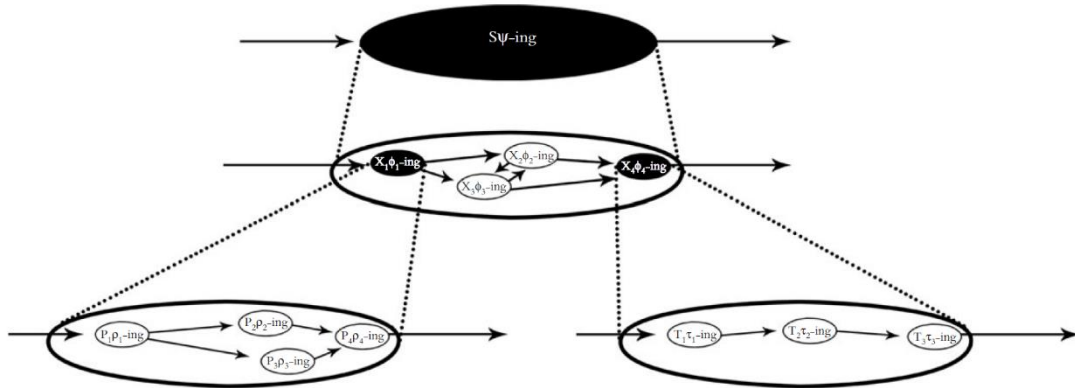


Figure 11: Levels are defined locally within decomposition hierarchies. Reproduced from Craver (2007, p. 194)

4.3.1. Framing the Explanandum: Defining Spatial Memory

Spatial memory is an excellent case for exploring a multilevel mechanistic explanation due to two interconnected reasons. Firstly, it is a cognitive phenomenon that has been extensively studied in the field of cognitive science. Secondly, due to the multilevel nature of cognitive science research,

investigations into spatial memory extend across various levels of organization. Research methodologies, findings, and theoretical concepts can range from higher psychological to lower neurophysiological and neurobiological levels. The identification of behaviors and their corresponding brain regions, functional circuits, neuronal populations, and subcellular mechanisms are all within the scope of relevant research plans.

Before embarking on a decompositional analysis to construct a model for a mechanistic explanation, it is crucial to establish an extensive definition of the explanandum phenomenon. What precisely are we seeking to explain? According to both Craver and Bechtel, this is an indispensable element of mechanistic decomposition. Craver (2007) stresses the importance of giving this step special attention, ensuring that the phenomenon is not illusory and that there exists a one-to-one correspondence to a mechanism. The latter means someone should not treat multiple phenomena as one (lumping error), or one phenomenon as if they are many (splitting error). In line with Craver's ideas, Bechtel (2008), in his methodology for mechanistic explanation, includes what he refers to as "phenomenal decomposition." This approach involves two types of decomposition: mechanistic, which consists of functional and structural decomposition (or, localization), and phenomenal decomposition, which is centered around defining the phenomenon. Bechtel argues that both forms of decomposition are necessary in order to fully analyze and explain a phenomenon. Hence, both Bechtel and Craver agree on the importance of exhaustively defining the phenomenon under examination.

We will examine all the above by exploring the phenomenon of spatial memory and, in particular, the case of a mouse navigating a Morris water maze. Spatial memory is a functional component of spatial cognition and encompasses two distinct phenomena (Aminoff & Daroff, 2003): 1) object manipulation and orientation, and 2) spatial orientation, which pertains to navigation, wayfinding, topographical orientation, and sense of direction in large-scale space. Although these two categories are related, they are mediated by different neural structures. The behavior of a mouse navigating in a Morris water maze falls under the second category, i.e., spatial orientation, which we will focus on in this discussion by first outlining the explanandum.

It's worth noting that the hierarchical organization of spatial cognition and memory is reflected in the relevant brain areas. Moreover, spatial cognition is not a unitary trait; it is a complex set of cognitive abilities that can function together or independently (Aminoff & Daroff, 2003) and, therefore, is explained by different mechanisms. For instance, spatial orientation can be encoded

using different reference frames, egocentric and allocentric³⁸, and each one, respectively, is mediated by different brain areas. Accordingly, navigation can be supported by two distinct types of mental representations: allocentric spatial orientation involves the formation of spatial mental maps, which takes place in the hippocampus, while the parietal cortex primarily mediates egocentric spatial orientation with the formation of mental routes, e.g., trajectories between a start and goal location, based on identifying specific landmarks in the environment (Fortin, 2008). The case that will be examined in this section, viz, the mouse navigating in a water maze, concerns allocentric spatial orientation.

We have already established an initial outline of what we are trying to explain, and part of that process involves localizing the specific brain areas involved. Essentially, we need to define and outline the phenomenon we are investigating while also recognizing the system responsible for producing it. This is reasonable because any phenomenon is the result of a system's behavior. On one hand, we define the phenomenon as the behavior of the system we are studying, and on the other hand, we determine the system's limits, including what is and isn't part of it. It is important to keep in mind that a system can exhibit multiple phenomena simultaneously (Kauffman, 1970). This means that a system can have many mechanisms, as each mechanism is for a specific phenomenon. Additionally, the same phenomenon can be realized by different systems, which is known as the multiple realization principle. However, it is important to note that simply breaking down the phenomenon without considering the system that exhibits it is not enough for a complete explanation. Mechanistic explanation considers both the system and the phenomenon it exhibits, setting it apart from functional analysis (Cummins, 1975), which is performed independently. It is only when both the system and the phenomenon are analyzed that a complete explanation can be achieved.

The subsequent stage in formulating an explanatory model involves uncovering the mechanism accountable for giving rise to the phenomenon in question, often referred to as the “explanandum phenomenon.” As articulated by Craver, a mechanism comprises systematically organized entities and activities identified at a level lower than the one where the explanandum phenomenon is primarily situated. These entities and activities collectively form the system and give rise to the

³⁸ Egocentric representations encode relations between landmarks in the environment and the organism, while allocentric representations depict only the relations among landmarks in the environment, without consideration of their relation to the organism (Fortin, 2008).

observed phenomenon. This constitutive relation extends hierarchically downward, with entities and activities constituting mechanisms that explain the behavior of entities at levels situated above them. This hierarchical arrangement comprises a series of levels related by parthood. Importantly, the concepts of levels and the components of mechanisms, namely, entities and activities, are interrelated and mutually defined. Both concepts are shaped by the experimental methodologies employed to delineate and elucidate the phenomenon under investigation. To illustrate this further, let us consider the case of spatial memory.

4.3.2. Levels of Analysis in Spatial Navigation

Navigating through space in an allocentric manner is a complex behavior that can be achieved using different strategies, each with a distinct neural basis. One such strategy involves the formation of a spatial map, which is controlled by the hippocampus. Evidence from various behavioral experiments and studies on hippocampal lesions supports this claim. The hippocampus is responsible for constructing and utilizing a cognitive map of the animal's environment, which allows it to navigate flexibly using different cues (Fortin, 2008). This cognitive map explains how a rodent can find its way to a specific location, as seen in experiments involving mice navigating through a Morris water maze.

In this context, we can proceed to our analysis focusing on two distinct levels of organization. The upper level, which I shall refer to as the *embodied/situated* level, is defined by the observable behavior of the animal, namely, the mouse's navigation through its interaction with the surrounding environment and by the animal per se. Beneath this level lies what I term the *brain areas'* level, housing the neural mechanisms responsible for the observed behavior. The explananda, relevant to these two levels, revolve around the following questions: 1) How does the mouse effectively navigate a water maze? The answer lies in the mechanism of constructing mental spatial maps, which facilitates spatial orientation. This leads us to the second question: 2) Where does the process of spatial map formation occur? This cognitive operation primarily unfolds within the hippocampus, in conjunction with several other brain areas. Each of these supplementary brain regions serves distinct functions and roles in the rodent's navigation. Specifically, the perirhinal, parahippocampal, and entorhinal regions function as links between complex representations housed in cortical association areas and the generation of mental maps within the hippocampus. Moreover, sensory systems engage in the processing and representation of environmental stimuli, while the prefrontal cortex provides the control and the navigation planning.

To apply Craver's methodology (2007) to the scenario presented above, we will proceed as follows: We can designate the mouse as the system, denoted as S , and define the act of navigating within its environment, as the system's behavior, specifically ψ -ing. Consequently, the mouse's navigation in a Morris water maze corresponds to $S \psi$ -ing. As previously mentioned, this behavior manifests at the upper level, referred to as the *embodied/situated* level. The underlying mechanism responsible for this phenomenon operates at a lower level, referred to as the *brain areas* level. This mechanism comprises spatially and temporally organized entities that act and interact, and in this context, these entities correspond to the various brain regions previously mentioned. Their respective roles, as outlined in the preceding paragraph, are the activities of the mechanism. Following Craver's terminology, we can identify the hippocampus, perirhinal, parahippocampal, and entorhinal regions, sensory systems, and the prefrontal cortex as the X s, representing entities. These entities, engage in various actions and interactions, hence, they are φ -ing. Together, they constitute the system S and its behavior, $(S) \psi$ -ing, which encapsulates the process by which the mouse navigates a particular environment.

4.3.3. From Behavior to Brain: Decomposing the Hippocampal Mechanism

Subsequently, our attention shifts to the hippocampus, the main contributor to this navigational behavior. Given that the requisite information for navigation is generated and stored within the hippocampus, the next explanatory inquiry delves into the hippocampal mechanisms, where two distinct processes unfold: the first facilitates the formation of spatial maps, while the second is responsible for the retention of this informational content. Regarding the first process, specific types of hippocampal neurons, originating from regions CA_1 and CA_3 , called *place cells*, fire at a high rate when the animal is in their place field, meaning, in a particular location within the environment. Their role lies in encoding multisensory stimuli generated as the animal occupies this specific spatial location. A collective assembly of these place cells collectively constructs a topographic map, effectively representing the environmental layout. The animal employs this encoded spatial information to compute navigational trajectories (Fortin, 2008).

The second mechanism, centered on information storage, is served by dedicated *trisynaptic neural circuits* that receive inputs from various brain regions. The central excitatory synaptic circuitry within the hippocampus consists of three functionally organized modules, as illustrated in *Figure 12*: the first module transmits information from the entorhinal cortex to the dentate gyrus of the hippocampus; the second conveys this information to area CA_3 ; and the third links area CA_3 to CA_1 .

After this last step of synapsing in CA₁, the information exits the hippocampus via projections from CA₁ pyramidal neurons. Subsequently, this information is relayed back to subcortical and cortical structures (Sweatt, 2007).

Hippocampus is a brain area and as such, its behavior pertains to the *brain areas* level, where the hippocampus assumes the role of system *X*, and its behavior (φ -ing) pertains to the construction of mental maps. But how is this behavior precisely executed? To address this question, I introduce a third level, situated beneath the previously delineated two levels (*embodied/situated* and *brain areas* levels). This level pertains to cellular components, encompassing cells and synapses, and therefore is termed the *cellular* level. At the cellular level, the component entities (*Ps*) within the mechanism are place cells and trisynaptic neural circuits. These entities are actively engaged in specific activities (ρ -ings), which encompass the detection of place fields and the storage of information, respectively.

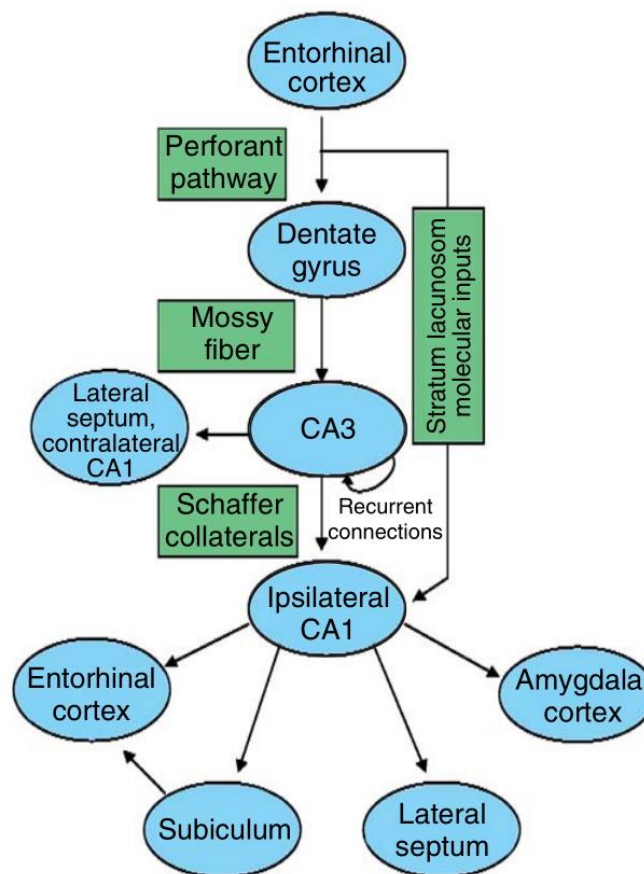


Figure 12: The entorhinal/hippocampal system. This panel diagrams the principal inputs, outputs, and intrinsic connections. Reproduced from Sweatt (2007, p. 299)

4.3.4. Cellular and Subcellular Mechanisms of Memory Formation

Let us delve even deeper into our exploration of the phenomenon under investigation. Now, our question shifts to understanding how all the encoded information from the previous steps is stored in memory. For this, let us first introduce the last (the fourth and lowest) level which will be referred to as *subcellular*. At this *subcellular* level, we encounter another crucial mechanism known as *Long Term Potentiation* (LTP), which plays a central role in synaptic plasticity. LTP is considered a molecular mechanism for high-order learning and memory in mammals, facilitating adaptations within the central nervous system, with a prominent presence in the hippocampus and other brain regions. The underlying idea supporting this mechanism harks back to Hebb's postulate and can be summarized as follows: "Memories are preserved by altering the strength of connections between neurons in the CNS" (Sweatt, 2007, p. 296). This concept was substantiated by the pioneering work of Bliss and Lomo (1973). Through electrophysiological experiments (stimulations and recordings) conducted on rabbit hippocampal synapses, they uncovered two key phenomena collectively termed LTP, which endured for extended periods (many hours). First, they observed a substantial strengthening of synaptic connections following a short burst of high-frequency (tetanic) stimulation. Second, when a constant synaptic input reached the cells, they exhibited a heightened likelihood of firing action potentials. These twin phenomena constitute the basis of synaptic plasticity.

The above description provides a brief overview of how information is stored through synaptic plasticity and the mechanism that explains this process. Here, we can perform a componential analysis, analogous to what was described in the preceding levels. We consider a neural circuit (P) as our system, with its primary function being information storage (ρ -ing). When we delve into the more detailed, *subcellular* level, we encounter the working components (Y s), which are synapses. These synapses engage in activities (σ -ings) involving interactions between presynaptic and postsynaptic cells, ultimately leading to the phenomenon known as LTP. Therefore, LTP serves as the mechanism through which the hippocampus stores information by reinforcing its cellular connections. This analysis could theoretically extend to the (even lower) *molecular* level, focusing on the molecular mechanisms responsible for the chemical and electrical activities of nerve cells, particularly concerning the activation of the NMDA receptor. However, we won't explore this mechanism further at this point since we have already illustrated the general idea of constitutive mechanistic explanation, which entails a recursive decomposition of entities and activities into even lower levels of organization.

Let us bring everything together using a schematic representation inspired by Craver (2007). In *Figure 13*, we have a hierarchy of mechanisms ranging from a higher *embodied/situated* level to a lower *subcellular* level. At each level, we talk about a system and its corresponding behavior, where behavior answers the question “what.” The mechanism that explains this behavior, addressing the “how” question, is presented at the level just below, with a set of entities engaged in specific activities. This organized collection of entities and activities constitutes and gives rise to the phenomenon we’re trying to explain. Now, when we want to go deeper and explain how a particular entity performs its activity, we follow a similar analytical process. In this case, the entity is the system, and the activity pertains to the specific behavior of the system that we aim to explain. So, just like in the previous stage and as shown in the figure, at the higher level, we have the system S performing ψ -ing, which is decomposed into its constituent entities X s and activities φ -ings at a lower level. This approach continues recursively as we want to explain how X s φ -ing happens, breaking it down into entities P s and activities ρ -ings, and to explain a P ρ -ing, we further decompose it into entities Y s and activities σ -ings, and so forth. In the context of the spatial memory case we discussed earlier, this schematic representation spans across four levels of mechanisms. However, it’s important to note that, as Craver points out, this division into levels is a simplification. Additional levels can be introduced between, above, or below the ones we’ve mentioned.

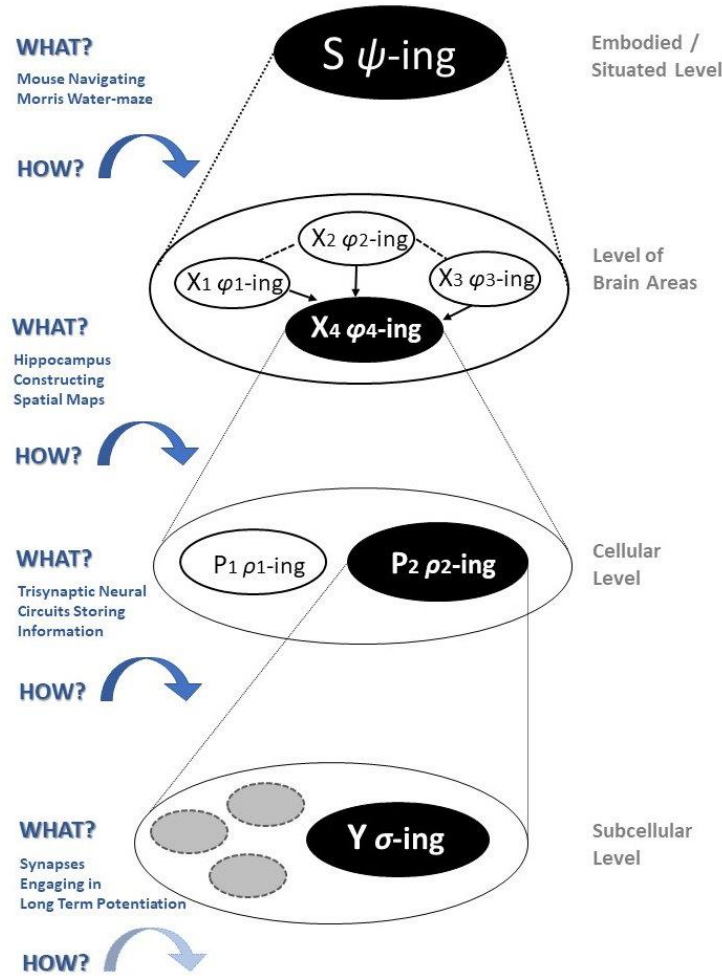


Figure 13: Four mechanistic levels for the explanation of the behavior of a mouse, navigating a Morris water-maze. 1). Level of spatial memory. S : mouse $\psi\text{-ing}$: navigating the water-maze. 2). Level of Brain Areas. Entities: X s, activities: $\varphi\text{-ings}$. X_1 : perirhinal, parahippocampal, and entorhinal regions, $\varphi_1\text{-ing}$: acting as an intermediary link. X_2 : sensory systems, $\varphi_2\text{-ing}$: processing and representing environmental stimuli. X_3 : prefrontal cortex, $\varphi_3\text{-ing}$: controlling and navigation planning. X_4 : hippocampus, $\varphi_4\text{-ing}$: constructing of mental maps. 3). Cellular Level. Entities: P s, activities: $\rho\text{-ings}$. P_1 : populations of place cells, $\rho_1\text{-ing}$: place field detection. P_2 : trisynaptic neural circuits, $\rho_2\text{-ing}$: information storing. 4). Subcellular Level. Entities: Y s, activities: $\sigma\text{-ings}$. Y : synapses, $\sigma\text{-ing}$: long term potentiation.

4.4. SUMMARY AND CONCLUSIONS

In this chapter I continued the preceding discussion about compositional explanations in biological phenomena, focusing now on the constitutive explanations that are structured according to the norms of the new-mechanistic account. I explored how it synthesizes and extends the concepts of levels and decomposition in explanation, which are the two pillars of the dissertation, extensively discussed in the first two chapters. By tracing the development of mechanistic explanation as it

evolved from the earlier incarnations of the idea in describing and explaining machines to its modern reformulation in a way that is more suited to dynamical and complex systems, I explored the core concepts and definitions that characterize the new-mechanistic approach. I then centered on Craver's account of constitutive mechanistic explanation in neuroscience. I gave special focus to his suggestion about levels of mechanisms that form the basis for the method of mechanistic decomposition he proposes. I then referred to the important distinction between etiological and constitutive explanations since this dissertation focuses on the latter. This discussion continues and specializes in the exploration of compositional explanations presented in Chapter 3, showing how constitutive mechanistic explanations provide a comprehensive framework for understanding multi-level biological systems.

A significant portion of this chapter was dedicated to exploring the scope of mechanistic explanation by examining the various interpretations of the concept and addressing its boundaries and challenges. This included examining what qualifies as a mechanism, addressing borderline cases, and the difficulties posed by highly complex biological systems. In this context, I considered the balance between model accuracy and idealization in mechanistic descriptions, extending the discussion of biological complexity from Chapter 2 to inquire how the new-mechanistic approach navigates the tension between decomposability and whole system properties.

The detailed case study on spatial memory explanation presented in the last part aimed to demonstrate the practical application of decomposing a phenomenon into a hierarchy of levels, based on Craver's (2007) norms for mechanistic explanation. This example illustrated how the new-mechanistic approach integrates the concepts of levels, decomposition, and compositional explanation discussed in previous chapters. It extended across multiple levels of biological organization, from the embodied/situated level where an animal's behavior is observed down to subcellular mechanisms, and revealed how constitutive mechanistic explanations can span from observable behaviors to underlying neural processes, bridging the gap between higher-level phenomena and lower-level mechanisms.

While this approach offers a powerful framework for explaining biological phenomena, questions remain about its applicability to mental phenomena. The next part of the dissertation will explore this issue in depth, examining whether and how mechanistic explanations can accommodate the unique, multilevel, and multifaceted nature of phenomena studied by cognitive science. This will involve considering how mechanistic levels of organization align with the various notions of levels

used within cognitive science, and whether the idea of decomposition in constitutive mechanistic explanation can be incorporated into cognitive models that explain mental phenomena across multiple levels of description and composition. As we move into this exploration in the second part of the dissertation, we will build on the foundations laid in this chapter, considering how the strengths of the new-mechanistic approach might be adapted or extended to address the specific challenges posed by mental phenomena.

PART II

EXPLORING THE APPLICABILITY OF MECHANISTIC EXPLANATIONS TO MENTAL PHENOMENA

In the second part of the present inquiry, I explore the applicability of the new-mechanistic framework in mental phenomena regarding explanations that proceed with decomposing the phenomena and their instantiating systems into hierarchies of levels (constitutive explanations). This exploration will address two aspects of the question: In the first chapter, I will explore whether mechanistic levels of organization, as described by Craver (2007), can accommodate the special, multilevel, and multifaceted study of mental phenomena, and in the second, how the idea of decomposition applies to cognitive phenomena and how can it be combined with mechanistic decomposition providing an integrated explanatory model that captures both information and physical processing aspects of the phenomenon and its implementing system. What is been explored is the possibility of an integrated cognitive-mechanistic model that could explain mental phenomena in multiple levels of description and composition, viz, all the way down.

CHAPTER FIVE – MULTIPLE LEVELS AND COGNITIVE SCIENCE EXPLANATIONS

5.0. OUTLINE

In the fifth chapter, I begin a more specialized treatment of the subjects examined in the first part, exploring whether the new-mechanistic explanation can be applied to phenomena studied by cognitive science, with a focus on the concept of levels. Here, I aim to examine the extent to which the concept of mechanistic levels is compatible with the various notions of levels encountered during the investigation of mental phenomena by the fields that constitute the multidisciplinary character of cognitive science. I begin with an introductory presentation of the particularities of mental phenomena instantiated in biological systems, emphasizing what differentiates them from mere biological phenomena and what the implications are for the models employed to explain them (5.1. *The Unique Nature of Mental Phenomena in Explanatory Frameworks*). I refer to intentionality and information processing as defining features of mental phenomena (5.1.1.) and highlight two important distinctions, the first concerning the role of information and representation in cognitive systems (5.1.2.) and the second concerning the vehicle vs content distinction as a key aspect of cognitive explanations (5.1.3.). In the second part of the chapter I examine the role of multiple perspectives in cognitive science, providing an overview of diverse notions of levels used in mental phenomena research (5.2. *Navigating the Levels: The Role of Multiple Perspectives in Cognitive Science*). I discuss levels of organization, processing, analysis, description, and realization (5.2.1.) and present visual perception as a model for multilevel explanations in cognitive science (5.2.2.). Next, I aim to inquire whether it is possible to combine cognitive science explanations with mechanistic ones (5.3. *Reconciling Cognitive Science Explanations with Mechanistic Levels*), and for this, I revisit Craver's account, examining his ideas on mechanistic levels and the potential to be applied to explaining mental phenomena (5.3.1.). The challenge of incorporating informational content into levels of mechanisms is explored in depth in the last section of the chapter (5.3.2.).

5.1. THE UNIQUE NATURE OF MENTAL PHENOMENA IN EXPLANATORY FRAMEWORKS

The previous chapter ended with a fully developed application of mechanistic decomposition, which incorporated Craver's norms of mechanistic explanation and his framework of mechanistic levels in the context of spatial map formation during mouse navigation in a water maze. The case study concluded by describing in increasing detail a hierarchy of neurobiological mechanisms relevant to the task. While this case study concerned a mental phenomenon - namely, the mouse's behavior - its constitutive explanation was exhausted by the neurobiological account; that is, it was limited to entities and processes concerning only the neurobiological substrate in which it is instantiated without explaining what happens to the content, informational and representational³⁹, that corresponds to these substrate transformations. However, it remains an open question whether this is sufficient for considering a mechanistic explanatory model as a fully integrated explanation within the domain of mental phenomena.

To fully explain a cognitive phenomenon occurring in a biological organism, it's essential to describe multiple aspects, not just the mechanisms related to its neurobiological substrate. Importantly, as a psychological phenomenon, it must also be explored from a psychological perspective, rather than solely as an outcome of specific neurobiological processes. Psychological phenomena are typically observed through behaviors, which is why we refer to operations occurring at the behavioral level. Cognitive psychology is the field that studies cognitive phenomena at the behavioral or psychological level. However, observed behavior and its triggering conditions alone are insufficient to fully explain cognitive phenomena. A comprehensive understanding of cognition requires knowledge of many additional factors. Cognitive Science, a multidisciplinary field focused on mental phenomena, is built on a fundamental principle: to open

³⁹ In Cognitive Science and Philosophy of Mind, the terms "informational content" and "representational content" are often used to distinguish different aspects of mental states or cognitive processes. While these terms are related, they refer to different concepts. Informational content refers to how a mental state or cognitive process carries information about the world. For example, in perception, a visual experience may have informational content about the presence of a red apple in the environment. Representational content refers to how a mental state represents or is about something and concerns the internal representation of an object or state of affairs within the mind. It is the intentional or semantic aspect of a mental state. More about this crucial distinction will be mentioned in the following sections.

the *black box* of the mind⁴⁰ and investigate mental operations and how the external world is represented within it. In practice, this means that a cognitive explanation goes beyond simply describing mental phenomena in terms of stimulus and observed behavior. Instead, it emphasizes how responses are generated through processes within the mind and brain. In essence, Cognitive Science aims to uncover the underlying mental mechanisms.

It becomes then apparent that while neurobiological mechanisms can be described in sufficiently great detail, purely neurobiological explanations such as the one presented in the previous chapter do not fully capture the mental aspect of the phenomenon since there is a need to account also for the psychological operations. To determine whether a comprehensive explanation of a cognitive phenomenon that satisfies the criteria of the new mechanistic explanatory framework is, after all, feasible, we first need to delve into cognitive phenomena. This will help us identify what gives them their distinct character and sets them apart from other biological phenomena. This will be the subject of the first section.

5.1.1. Intentionality and Information Processing: Defining Features of Mental Phenomena

What makes mental phenomena unique? One of the key aspects differentiating mental phenomena from other biological phenomena, such as the functioning of the kidneys or the execution of a movement by the musculoskeletal system, is *intentionality*. Intentionality refers to the property of the mind to be directed toward its objects. This idea, first proposed by philosopher Franz Brentano in his influential work “Psychology from an Empirical Standpoint” (1874), has been widely accepted as a defining characteristic of mental phenomena. Intentionality is often described as the mental state of being about something. It is the quality that allows the mind to represent the world and form beliefs and desires about it. Essentially, it is the property that allows the mind to be directed toward objects and events in the external world and to relate to them in a meaningful way. In this sense, intentionality is a key feature of mental phenomena that sets them apart from other biological processes. While other biological systems may have functions or goals, they do not have the same capacity for directedness and meaning-making as the mind. This characteristic of

⁴⁰ This metaphor is about the limitations in understanding the inner workings of the human mind and implies that the mind operates like a sealed container, where information goes in and behavior comes out, but the mechanisms responsible for the transformation of inputs into outputs remain hidden from direct observation. The origin of this metaphor can be traced back to behaviorism.

intentionality is what makes mental phenomena unique and deserving special consideration in the study of the mind and cognition. As Brentano (1874) states:

Every mental phenomenon includes something as object within itself, although they do not do so in the same way. In presentation, something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired, and so on.

This intentional inexistence is characteristic exclusively of mental phenomena. No physical phenomenon exhibit anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves. (p. 68)

Mental phenomena are a distinct category of phenomena characterized by the fact that they involve the representation and processing of informational content within the mind. This has significant implications for the study of mechanisms, as it means that in addition to physical processes, such as the transformation of material substances, the explanation of mental mechanisms must also account for mental operations involving the transformation of symbolic representations⁴¹. In other words, while the mechanistic explanation of most biological phenomena can be understood in terms of physical processes and entities alone, the explanation of mental phenomena requires consideration of both physical processes and mental operations and entities (Bechtel, 2008).

The distinct way needed to study systems that possess intentionality is reflected in the explanations provided for these systems and the phenomena they instantiate. This can be easily observed if we compare two explanations concerning biological phenomena, the second of which is a cognitive one. Let us first recall the mechanism by which the muscular system functions⁴², which is not intended to represent and process information but rather to produce movement at a joint: The motor neuron enters the muscle and conveys a brain-generated motor pattern through the motor

⁴¹ Symbolic transformations refer to the manipulation of symbolic representations within the mind. These representations can be thought of as mental symbols that stand in for real-world objects, concepts, or ideas. The manipulation of these symbols involves processes such as combining, rearranging, or abstracting them in various ways. For example, the mental operation of language comprehension involves the symbolic transformation of words into mental representations of their meanings, which can then be combined and manipulated to form higher-level representations of the overall meaning of a sentence or paragraph. In the visual system, mental symbols are features of seen objects that through stages of processing, are combined into a complete representation and recognized as a specific object.

⁴² Fully developed in 3.3.

plates. This motor command initiates a contraction process where the reduction of the muscle's length results in a traction force on the two muscle tendons, which transfer this force to the connected bones. As soon as the brain sends the electrical stimulation pattern through the spinal cord to the motor units, the stimulation spreads to the muscle fibers, where calcium is released through the sarcolemma. Calcium initiates muscle contraction by activating a molecular mechanism inside the sarcomeres called the cross-bridge cycle. Cross-bridge cycles are continuously formed throughout the contraction, and they cause a sliding movement between the sarcomere's actin and myosin filaments, which are partially overlapping. The sliding of these filaments leads to the contraction of the sarcomeres, and the contraction of many sarcomeres results in the contraction of the myofibrils that contain them.

We will contrast this description of the muscular mechanism with a neurobiological explanation concerning a perceptual phenomenon executed by the visual system. Visual processing begins with sensory stimuli (light) transduction into nerve impulses in the retina's photoreceptors. The mechanism of generating receptor potential involves the transformation of electromagnetic radiation (light) into changes in the membrane permeability of photoreceptors in the retina (rods and cones). This process involves the modification of ion diffusion through the membrane, which in turn changes the transmembrane potential of the receptor. If this potential exceeds a certain threshold, it will trigger the generation of action potentials that propagate along the receptor's nerve fiber. This neurobiological mechanism of receptor potential⁴³ consists of a series of physical transformations that convert sensory stimuli into electrical signals that can be transmitted and processed by the nervous system. The stimulation of multiple receptors generates action potentials that form specific patterns of activation, which are then transmitted and subject to various transformations as they progress through subsequent stages of the visual pathway.

It is now time to compare the two cases described. Essentially, both involve generating and transmitting an activation pattern through a neuron or group of neurons. This shared characteristic is the reason they were selected for comparison. In the case of the muscular system, the activation pattern is transmitted to a group of muscle fibers, causing them to produce force. This force is then exerted on the bone connected to the muscle and produces movement with specific characteristics (such as intensity, duration, speed, and rate). The whole story ends at this point: the nerve signal

⁴³ Described in detail by Guyton & Hall (2006)

transmitted to the muscle is translated into a muscle signal and the corresponding movement pattern. The information conveyed through the initial activation pattern serves a specific purpose: movement production. But what about the information conveyed through an activation pattern in a mental phenomenon? In the case of visual perception, the information itself is the purpose of the system's operation; specifically, the visual system's goal is the internal representation and understanding of the perceptual object. This highlights a crucial distinction between the two systems we're examining, specifically regarding their operational purposes. In the muscular system, information serves as a means to an end. In contrast, for the mind, processing information is the end itself – it's the very purpose of its operation.

For representing internally and understanding a specific visual scene, information processing is needed, which relates to another significant difference between the two cases: cognitive systems are engaged in both physical and informational processing. Indeed, in visual perception, two distinct processes occur: the physical transformations of electrical signals in the visual pathway occur concurrently with transformations of the informational content carried by the activation patterns formed by the stimulation of multiple receptors. These patterns carry information about the sensory stimuli being processed, and this information is subject to various transformations as it progresses through subsequent stages of the visual pathway. The informational content carried by an activation pattern might represent the visual features of an object being viewed, such as its shape, color, and spatial location. As this information is transmitted and transformed through the visual system, it may be combined with other information about the object or used to generate higher-level representations of the object's meaning or significance that contribute to the generation of meaningful representations of the external world within the mind.

This double-aspect function of the visual system reflects what Hubel (1995, p. 2), in his influential work on the neural basis of vision, briefly describes:

The questions that I will be addressing can be simply stated. When we look at the outside world, the primary event is that light is focused on an array of 125 million receptors in the retina of each eye. The receptors, called rods and cones, are nerve cells specialized to emit electrical signals when light hits them. The task of the rest of the retina and of the brain proper is to make sense of these signals, to extract information that is biologically useful to us.

Hence, the visual system simultaneously performs two types of processes: physical and informational. Electromagnetic radiation, or light, carries information that is decoded by the visual system through the process of light transduction. This process occurs in a series of stages, beginning at the retinas and progressing through the lateral geniculate bodies to the primary (striate) visual cortex and other higher visual areas located in the occipital, parietal, and temporal lobes. At each stage of the visual pathway, outputs, or action potentials, are transmitted to the cells of the next stage, which are activated or inhibited based on the specific features of their receptive fields. These cells respond to different aspects of the visual scene, and as the processing progresses to higher stages, the behavior of the cells becomes more complex as they respond to more intricate input patterns collected from multiple visual areas with a wide range of activation features. In the early stages of the visual processing pathway, the encoded information is relatively simple, consisting of basic distinctions such as the presence or absence of light in different areas. This information is transmitted to the next stage, where it is collected and used to activate more complex patterns that correspond to more complex information, such as lines and edges. As the processing progresses to higher stages, the complexity of the encoded information increases along with the complexity of the cells' response. As Hubel (1995, p.2) states, "The result is the scene as we perceive it, with all its intricacy of form, depth, movement, color, and texture."

Mechanism descriptions like the two summarized above are common in textbooks of related fields, as illustrated by the following passages. The first passage describes a mechanism for the musculoskeletal phenomenon, while the second regards the visual perception phenomenon:

Contraction is initiated by an electrical stimulus from the associated motor neuron, causing depolarization of the muscle fiber. When the fiber is depolarized, calcium is released into the cell and binds with the regulating protein troponin. The combination of calcium with troponin acts as a trigger, causing actin to bind with myosin, beginning the contraction. (Oatis, 2009, p. 47)

... the eye is not a mere camera, but instead contains sophisticated retinal circuitry that decomposes the retinal image into signals representing contrast and movement. These data are conveyed through the optic nerve to the primary visual cortex, which uses this information to analyze the shape of objects. It first identifies the boundaries of objects, represented by numerous short line segments, each with specific orientation. The cortex

then integrates this information into a representation of specific objects, a process referred to as *contour integration*. (Kandell, Koester, Mack, & Siegelbaum, 2021, p. 545)

When observing the terminology used in these two passages and the previous descriptions of the related mechanisms, a profound difference between the musculoskeletal and the perceptual phenomenon is revealed. This difference could be insightful concerning how an integrated explanatory model should treat these phenomena. First, let us examine the musculoskeletal phenomenon. If we compile a list of the terms used to describe it, we will find words like *muscles, tendons, brain, spinal cord, motor neuron, troponin, bones, motor units, muscle fibers, motor plates, calcium, sarcomeres*, etc. All the above are terms that correspond to physical entities that are processed by the musculoskeletal system. What about the perceptual phenomenon? Here, the list seems to have more variety. Along with the terms describing physical entities such as the *eye, optic nerve, primary visual cortex, lateral geniculate nucleus, retinal circuitry, neurons*, and the like, we also encounter terms that describe entities that belong to the features of the visual scene, namely, entities of the external world that are represented internally: *image, objects, movement, shape, boundaries, orientation, contrast, contour, lines, edges, line segments* and so on. This difference reflects the duality in studying mental phenomena, concerning their neurobiological and representational nature. In what follows, I will further explore the relation between information and representation in shaping mental content.

5.1.2. Information and Representation in Cognitive Systems

So, the two fundamental characteristics of the mental realm are that mental phenomena involve information processing and intentionality, the latter being closely linked to the representational content of a mental state. The role of information in cognitive systems pertains to how a mental state covaries with or is caused by states of the world. It is often understood in terms of causal or nomological (law-like) relations between the mental state and the environment. For example, in the visual system, the presence of a stimulus that moves in a specific direction is nomologically connected with the firing of direction-selective retinal ganglion cells when the stimulus is projected onto their receptive field.

A specific event of the external world carries information that the mental state has as its informational content. Representational content, on the other hand, refers to what a mental state is about or what it represents from the subject's perspective. It is the intentional or semantic aspect of a mental state and concerns how the mind interprets or uses the information it receives. Possessing

informational and representational content presents significant differences, although there have been attempts to reduce the latter to the former. In this regard, some theories about the content of mental states focus more on information, known as informational theories of content, while others focus on representation (representational theories of content). To understand the difference, we must first examine how the concept of information is used in various contexts and how it is applied in cognitive science.

The concept of information is widely prevalent. We certainly encounter it in cognitive systems, biological systems, and in numerous other fields in various forms. To mention a few examples, in Computer Science and Information Theory, we encounter Shannon information, referring to the quantitative measure of uncertainty reduction; algorithmic information, which is about the complexity of the shortest program that can produce a given string; and quantum information, that is, information stored in quantum systems. In Physics, the most common use of information relates to entropy and the organization of physical systems (thermodynamic information). In Biology, several notions of information are used, such as genetic information encoded in DNA and RNA sequences, epigenetic information regarding heritable changes in gene expression not involving changes to DNA sequence, and in neurobiology, we encounter neural information concerning signals transmitted through neural networks. The concept appears quite differentiated in cognitive science and psychology, where we encounter it as semantic information related to notions such as meaning and mental content. Let us examine how various conceptualizations of information differ.

As extensively argued in Davies & Gregersen (2014), which discusses the role of information in our understanding of reality, information is a fundamental part of the universe, perhaps more fundamental than matter or energy. Focusing on the biological domain, the role of information is tightly connected with biological functions. John Maynard Smith (2014) argues that biological sciences are inherently informational and Terrence Deacon (2014) as well as Bernd-Olaf Küppers (2014) discuss the emergence of semantic levels of information in biology. Specifically, Küppers states that "all elementary processes of life are governed by information" (Küppers, 2014, p. 170), suggesting that information is not just an aspect of cognitive systems, but a fundamental feature of all living systems. He adds that information processing and understanding information are distinct concepts - the latter pertains to the semantic aspect of information. The author argues that understanding information requires background information of at least the same complexity. This suggests that cognitive systems must have a certain level of complexity to process and understand complex information from their environment.

Fred Adams, (2016), ponders what makes mental systems different from anything else that can function as an information carrier or processor: "What can cognitive creatures do that the thermometer cannot do?" (p. 324). The author argues that if information processing (and the resulting behavior) were sufficient to characterize a system as mental, then the entire universe would be a mental system. As he writes: "If being a conduit of information were sufficient for cognition, the whole universe would think" (Adams, 2016, p. 323). However, cognitive systems are not only information processing systems. They can represent things and have concepts and understanding, which they acquire when their mind processes the received information by building concepts. They do not simply record information; they transform it into beliefs and thoughts. Additionally, they can misrepresent things, whereas information cannot be misrepresented. "It cannot because information is carried by the lawlike connection between properties in nature" (p. 326). It is what it is. "It is *nomically tied to the truth*" (p. 325). Adams mentions an example regarding smoke and fire: smoke provides information that there is fire, and this is a lawlike connection between smoke and fire. However, someone can falsely believe that there is smoke and think that there is fire. Believing and thinking are representations. Furthermore, an information processing system (like the thermometer that Adams uses as an example) cannot plan far into the future; it is tied to the *here and now*. Finally, a mind represents something in a specific form (the vehicle of the representation). Information has no form.

Given these considerations, Adams concludes that "cognition is different from mere information processing" (p. 325). He adds that "cognitive systems are different in that they utilize representational vehicles (symbols, concepts) that represent or have meaning." (p. 325). What is the difference according to Adams, after all? "Cognitive systems process information and generate symbols with meaning. Noncognitive systems do not. They remain forever locked in the present, processing information but not meaning" (p. 328). Nevertheless, Adams states that "there is an informational origin to all concepts and meaning" (p. 328).

Taking all these into account, we can argue that although there are biological and other systems that use information, the concept of information in the cognitive context is very different. Therefore, when we use it, we should keep all of the above in mind. So, what sets them apart when it is stated that mental systems or cognitive agents are distinguished from biological systems or other information-processing systems as primarily mental? First, it is the different use of the concept of information. Second, what differentiates them is their representational character. As we have seen, this representational character may be incorporated to some degree in the concept of

information, either by considering that it is grounded in it or that it has a different and richer/more complex form compared to mere information.

Some philosophers and cognitive scientists have proposed theories that blur the lines between information processing and representing, or that treat them as closely related. More extreme views fall under eliminative materialism or considerations of mental content within the scope of enactive or extended approaches to cognition, which deny that there are representational contents in the mind. It should be noted, however, that in most cases, none of these thinkers explicitly argue that information processing and representational content do not differ significantly. Rather, their work sometimes suggests closer relations between these concepts than traditional views might hold. For example, Paul Churchland as a proponent of eliminative materialism argues that our folk psychological concepts (including many representational notions) will eventually be replaced by neuroscientific explanations (Churchland P. M., 1981). However, he does not reject the informational character of mental events, so we can infer that if he accepts any content, it is the informational content that can be reduced to neurobiological states with which it is systematically correlated.

A milder position is that of Ruth Millikan (1989) who in her biosemantics theory emphasizes that representations should be considered based on the biological functions they serve, and are grounded in information. However, she provides important aspects that differentiate human representations from those of bacteria and other simpler biological organisms. These aspects include the ability to store representations, the existence of mental states with representational content that are not tied directly to actions, the capacity to combine representations to form new beliefs (i.e., mental states with representational content), and other distinguishing features. This approach is more akin to the view that information is very closely related and allied to the concept of representation, but there are significant points that differentiate them, as mentioned by Adams earlier.

Fred Dretske's work also suggests a close relationship between information and representation. While he doesn't equate informational with representational content, his informational semantics theory proposes that mental representations derive their content from the information they carry (Dretske, 1988). In his informational theory of content attempted to ground representational content in informational content (Dretske, 1981) and to bridge philosophy and cognitive science by providing an information-based account of knowledge, perception, and belief. Through his account,

he aimed to explain the intentional properties of mental states in terms of information-theoretic concepts. He defined knowledge as information-caused (or causally sustained) belief, beliefs as internal representations that acquire meaning through their information-carrying role and developed through learning, and, lastly, perception as a stage in the processing of sensory information in which information about a perceptual object is coded in an analog form from where a relevant belief may be generated (Dretske, 1983). Due to its information-theoretic approach to explaining mental phenomena, grounded in objective, physical processes, Dretske's theory is often used to interpret mental content in theories that attempt to combine neurobiological and psychological explanations and provide a framework for understanding mental representations and content in terms of information processing, thereby offering a potential bridge between more traditional cognitive theories and information-based approaches to mind and cognition. One such example is Boone and Piccinini's (2016) account, which will be examined in the next chapter.

Concluding this discussion regarding the concepts of information processing and intentionality (possessing representational content), and having seen to what extent these two are connected, we should remind ourselves of what was emphasized in the previous section: cognitive systems as information processors have a very important difference from other systems that carry or process information: for them, information is the purpose. They represent it internally, modify it, combine it, create concepts with it that they understand, form true or false beliefs about how the world is, imagine various hypothetical scenarios through them, and so on. No other biological system has the processing and representation of information as an end in itself.

All of the above suggests that mental systems process information in a very special and unique way. And therefore, the concept of information used in them is also very different and much richer than that used in other systems, biological or not. Information processing is used by such systems to construct rich representations that are understood, perceived, believed, and used in reasoning and behavior production. This conceptualization of information we will use here when referring to *information processing systems* and *informational content*. After all, in the examples of visual perception used here, the mental processing we are studying and trying to explain does not only concern the information of the presence of a specific perceptual object in front of an observer but the building of its representation, with all the richness it can have as the content of a perceptual state. This is the content that an explanatory endeavor should try to incorporate into an explanatory model, and which, proposals that easily provide a solution to the application of mechanistic explanations to mental phenomena, often undermine. Let us now delve a little further

into another important distinction concerning the study and explanation of mental phenomena. This concerns the vehicle and the content of a representation.

5.1.3. The Vehicle-Content Distinction as a Key-aspect of Cognitive Explanations

In the philosophy of mind, the uniqueness of mental phenomena relates to the conceptual distinction between representational content and representational vehicle (Millikan, 1991; Dennett, 1991)⁴⁴. Every mental event has two aspects that can be approached by two questions: What is represented in the mind? This is the content. How is it represented? This is the vehicle. The distinction between content and vehicle can also be applied to information, allowing us to speak of informational content and informational vehicle, albeit within a broader scope and context (not necessarily cognitive) for the latter. The informational content of a mental state concerns what the information is about or what it conveys, while the vehicle refers to the physical medium or format in which the representation is embodied or carried. When it comes to the physical vehicle of information, it can be understood as neurobiological configurations, chemical structures, electrical signals, and similar physical manifestations. However, information isn't limited to physical implementations alone. It can also be implemented in abstract structures, such as mathematical and logical constructs. Furthermore, it can be conceptualized as the state of a computational system or as a statement in a programming language. This abstraction allows us to discuss and manipulate information independently of specific physical implementations.

The idea of the vehicle and content of a mental state is commonly employed to differentiate between the mode of representation or experience of a mental state, such as language or imagery, and the substantive aspect of the mental state, including beliefs about the world or emotions in relation to specific occurrences. The mode of a representation concerns the role that it can play in a representational system and refers to the format of a representation which is the structure of the representational vehicle which is the bearer of the content (Block, 2023). As Block notes: "A single format can express different contents in different representational systems. To take a common

⁴⁴ The concept of the vehicle and content of a mental state can be traced back to the ancient Greek philosopher Aristotle (350BC/1984), who distinguished between the form and matter of a mental state. In his work "On the Soul" he argues that the form of a mental state is the way it is represented or experienced, while the matter is the subject matter or content of the mental state. Despite the fact that Aristotle's ideas have been interpreted and debated by philosophers over the centuries and his views on this topic have been revised and refined by later philosophers and psychologists, his work is considered to be an important foundation for this conceptual distinction.

example, one pronunciation of 'Empedocles leaped' in English is the same as one pronunciation of 'Empedocles liebt' in German, though they have different contents, as Donald Davidson noted (1968, p. 135) (p. 173)." He also provides examples of how the same content (e.g., "kittens are cute") can be expressed in different formats across languages.

The distinction between vehicle and content has important implications for studying mental phenomena, as it allows researchers to study the different aspects of mental states separately and to understand how they interact. In this way, researchers can study the neural basis of language, which is a vehicle for a mental state, without necessarily considering the content of the mental state that has language as its vehicle. Similarly, in the case of the visual system that we examined, the list of the physical entities concerns the neurobiological substrates of the vehicle, in this case, vision, while the second list includes the content of visual percepts.

This dual character of mental phenomena tightly related to the mind's function prescribes how mental phenomena should be explanatorily approached and modeled. As Dennett (1991, p. 147) points out, "... the general principle of the content/vehicle distinction is relevant to information-processing models of the brain in ways that have seldom been appreciated. In general, we must distinguish features of representings from the features of representeds." This necessity emphasized by Dennett results from the particular information processing and representational character of mental phenomena. Studying them while considering that every mental phenomenon exhibits a physical and an informational/representational aspect is necessary to capture the essence of mentality. The two-sidedness of mental events is further accentuated when the computer metaphor is taken into account. This metaphor, which was a major inspiration for understanding the nature of the mind, continues to inform most explanatory frameworks in cognitive science. According to this metaphor, the mind is an information-processing machine that manipulates symbols. The mind is responsible for representing the external world, and a crucial aspect of representations is that they are symbols that are manipulated by the brain. Representations possess both syntactic and semantic properties, with the former referring to the vehicles through which representations are conveyed and the latter referring to the contents of a thought (or any other mental event that involves representation). The syntactic properties of a representation determine its functional role within an agent's psychology, as the physical properties of a representation determine the causal relationships it can enter into with other representations (Rowlands, 2003). Therefore, both aspects of a mental phenomenon, along with the way they interact, must be studied.

In the same spirit, Bechtel (2008) points out in his account of mental mechanisms that the content and the vehicle of representations play a key role in understanding the function of the mind and especially emphasizes the importance of their interaction. He asserts that “what is critical to understanding an information processing mechanism is how content is changed as a result of causal operations that change vehicles, for it is in virtue of content that the mechanism is linked to the states outside the mechanism” (Bechtel, 2008, p. 25). Given that the mind operates as a means of interacting with and perceiving the external environment, the content of a representation serves as the link between the internal and external world. As a result, a comprehensive explanation of any mental phenomenon must involve parallel descriptions of both physical and informational mechanisms and an examination of the relationship between them at each level resulting from the recursive decomposition of the phenomenon and its instantiating system. A comprehensive understanding of cognition requires integrating and correlating findings from both informational and neurobiological processes at various levels.

5.2. NAVIGATING THE LEVELS: THE ROLE OF MULTIPLE PERSPECTIVES IN COGNITIVE SCIENCE

According to Carver (2007), to comprehend neuroscience, one must understand the type of explanations it provides (descriptive project) and this is the first step to distinguishing between good and bad explanations (normative project). This statement addresses a hypothetical question that would naturally arise: Why do we need a framework for explanations in the first place? Carver's perspective likely applies equally to cognitive (neuro)science, where the inquiry becomes even more pressing since, in this specific research domain, efforts are being made to integrate various explanatory approaches originating from diverse research fields, scientific disciplines, and methodologies. Could possibly the new-mechanistic account for multilevel (constitutive) explanations provide a way to integrate these approaches that comprise cognitive science? As this dissertation primarily examines two fundamental elements of mechanistic explanation —namely, levels and decomposition— we will commence our exploration here by focusing on the former, specifically, levels, and will pose the question: Can the new-mechanistic conception of levels capture the role that levels play in studying cognitive phenomena?

To answer this question, I will first explore the various notions of levels found in the theoretical constructs of cognitive science and the methodologies employed to study cognitive phenomena. I

will then examine how they are combined in a case study related to visual perception. As a science born from the collaboration and desire to integrate various disciplines, cognitive science aims to combine different levels of organization, analysis, and description, both methodologically and theoretically. These diverse conceptions of levels need to be clarified to enable an assessment of the applicability of the mechanistic explanatory framework to the explanations provided by cognitive science.

5.2.1. An Overview of the Diverse Notions of Levels in Mental Phenomena Research

Does cognitive science research program operate in a multilevel manner, targeting systems at various levels of organization, analyzing phenomena through diverse approaches, and developing theoretical constructs to describe and explain phenomena across a hierarchy of levels? A general overview of the field would evoke an affirmative answer. However, what precisely does this affirmative answer mean? In the first place, within the cognitive science field which is comprised of many different sciences, it is common practice, when studying cognitive phenomena, to investigate research questions by examining multiple levels of organization, that is, levels defined through how a system is composed, targeting entities at various scales, and multiple analytical viewing angles, which refer to how a system is studied. This multilevel approach aligns with the field's multidisciplinary nature, drawing on diverse areas such as psychology, philosophy, neuroscience, computer science/artificial intelligence, linguistics, and anthropology⁴⁵. Multidisciplinarity allows cognitive scientists to gain a comprehensive understanding of the mind and its functions by trying to integrate insights and methods from different research areas.

The multilevel approach of the cognitive science research program is evident in the diverse research methodologies employed by its constituent disciplines, which address various levels of organization and analysis. To illustrate this diversity without providing an exhaustive list, consider the following examples. Psychology uses both experimental and observational techniques to study organism behavior. Psychophysics combines physical and psychological properties to examine perceptual phenomena at the sensory system level. Brain imaging studies investigate neurobiological properties at the whole-brain level, while lesion experiments target specific brain

⁴⁵ In its early stages, Cognitive Science was envisioned as an interdisciplinary field that would emerge from the collaboration of six main scientific disciplines: cognitive psychology, neuroscience, artificial intelligence, linguistics, philosophy, and anthropology. These fields were considered essential in shaping the initial goals and framework of cognitive science as it was first conceived (Núñez, et al., 2019).

regions. Systems neuroscience employs models and techniques to address mental phenomena at the neuronal network level, and electrophysiology examines processes at the cellular level. The primary goal of combining these methodologies is to provide more comprehensive and unified explanations. This is achieved by integrating findings that support hypotheses and theories across multiple levels of analysis and organization. By doing so, cognitive science aims to create a more holistic understanding of mental processes and behavior.

To illustrate more practically the multilevel character in methodologies employed and research questions posed to explain a specific cognitive phenomenon, let us consider the example of body ownership perception within the research domain of multisensory perception. This field is notably multidisciplinary and methodologically pluralistic, as we will explore in what follows. In this multisensory phenomenon, impressions from different sensory modalities are combined to form a cohesive perception of a single owned body part (Ehrsson, 2020). How is such multisensory integration feasible? The objective is to identify the processes that mediate this perceptual phenomenon. Explaining its emergence necessitates decomposing the inquiry into smaller research questions. These come in two categories: questions specific to the phenomenon and those pertaining to the broader field of multisensory perception, applicable to the specific and other multisensory perceptual phenomena.

Let us examine some of these questions. Firstly, it is reasonable to inquire which perceptual modalities contribute to the emergence of the phenomenon. With a prominent experimental example being the use of the Rubber Hand Illusion⁴⁶ (Botvinick & Cohen, 1998), multiple studies have revealed that the main sensory modalities involved are vision, touch, and proprioception (Ehrsson, 2020). Addressing this question primarily relies on behavioral experimental research tools. A complementary question is how these sensory modalities cooperate to produce the phenomenon. This question can be broken down into two sub-questions. First, what perceptual rules govern the phenomenon? Answering this sub-question requires using behavioral experimental research tools and has uncovered the involvement of four rules: the temporal rule,

⁴⁶ The rubber hand illusion (RHI) is a well-known experimental setup widely used in neuroscience and psychology research to investigate the plasticity of body ownership perception. It involves creating an illusion where participants perceive a rubber hand as part of their own body as the experimenters simultaneously stimulate both the participant's real hand (which is hidden from view) and the rubber hand.

spatial rule, tactile congruence rule, and humanoid shape rule⁴⁷. The second sub-question aims to identify the types of interactions that occur: whether they are unimodal, multimodal, or crossmodal. This research question can be explored at the behavioral level, but other methodologies, such as neuroimaging studies, can contribute by identifying brain regions where visual and tactile signals integrate. Indeed, fMRI research has identified regions in the premotor and intraparietal cortex that respond to spatially and temporally congruent multisensory stimulation from the body and surrounding space (Ehrsson, 2020). This provides valuable clues regarding the presence and location of multisensory neurons in the brain, a topic further investigated through electrophysiological measures and single-cell data from nonhuman primates, aiming to uncover the neural processes involved. I will not delve further into this example, as I assume, the description provided is enough to demonstrate the multifaceted and multilevel nature of investigating cognitive phenomena.

This brief example demonstrates that in the field of Cognitive Science, various concepts of levels are employed when attempting to explain cognitive phenomena, that is, the concept of 'levels' does not refer to a single idea, but encompasses a variety of distinct notions. These notions of levels found in cognitive science, are more abstract and conceptual, offering various perspectives that assist researchers in examining cognitive phenomena from diverse angles and hierarchical viewpoints. Typically, in the biological sciences, when something is said to be at a higher level than something else, it often pertains to levels of organization. However, as we strive to delve deeply into biological cognitive phenomena and observe the scientific efforts for identifying the physical and cognitive processes that underlie them, we encounter different notions of level hierarchies. It is, therefore, useful to identify these levels' notions to gain a better understanding of the multilevel nature of cognitive science and how different levels contribute to shaping explanations.

⁴⁷ According to the principles of temporal and spatial congruency, which govern multisensory integration, when signals from different sensory modalities occur simultaneously (temporal congruency principle) and in the same location (spatial congruency principle), they will be integrated, leading to the perception of a unified multisensory experience. Regarding body ownership perception, the temporal rule emphasizes that our sense of ownership of a limb is contingent upon the synchronized timing of multisensory cues originating from that limb. The spatial rule pertains to the alignment of sensory signals in space. The tactile congruence rule addresses congruencies in various stimulus properties, such as the alignment of tactile sensations with visually perceived touches. The humanoid shape rule suggests that our perception of a limb as part of our body hinges on the congruency of the observed limb's shape and spatial configuration with that of the limb we feel (Ehrsson, 2020)

Next, I will take as a starting point the distinction made by Churchland and Sejnowski (1992) regarding various ideas of levels' conceptions, which pertain to discussions concerning the nature of psychological/cognitive phenomena and their neurobiological basis. I will discuss the three notions they distinguish: Levels of organization, processing, and analysis. Pylyshyn (1984) introduced levels of description, which is a notion associated with cognitive science research, and I will present this as related to levels of analysis. I will explore lastly, levels of realization, a concept typically encountered in philosophy of mind and related to levels of analysis, as well.

I will begin with the most common notion of levels concerning biological organization or organization within a biological organism: *levels of organization*. For Churchland and Sejnowski, they "are essentially anatomical, and refer to a hierarchy of components and to structures comprising these components" (1992, p. 18). As mentioned earlier, very often, when someone refers to levels, they mean levels of organization. In the context of Systems Theory, "the contemporary use of the term 'hierarchy of levels' refers to systems structured in layers or levels that have asymmetric relations" (Wu, 2013, p. 285). A very common asymmetric relation is that of composition, which can only be applied in one direction. In other words, we can say that lower levels compose higher ones, but not the other way around. We are referring here to nested levels in compositional hierarchies, where the lower levels have part-whole relations with the higher levels, and the hierarchy of levels emerges from the repeated application of decomposition at each lower level.

In the context of biological systems, multiple criteria determine the hierarchical classification of organizational levels. These criteria include the size of an entity, the magnitude of a phenomenon, the interactivity between entities, the duration of processes, and the spatial containment of lower levels within higher ones. For example, examining various methodologies used within the cognitive science research field, brain scans are considered to target a higher level of biological organization compared to single-cell recordings because the brain is larger (criterion: size of an entity), it includes neural cells as constituent entities (criterion: spatial containment of lower levels within the higher), and it can be investigated using different methodological tools (criterion: interactivity between entities). Another example comes from the study of cognitive processes (Newell, 1994): At the neural circuits level, typical operation durations are approximately 10 milliseconds, while at the lower neuronal level, durations are on the order of 1 millisecond (criterion: duration of processes). Churchland and Sejnowski's (1992) levels of investigation, which pertain to entities and their associated processes at various scales, serve as a reference point in the field of Cognitive Science.

They identify 7 levels of neurobiological organization primarily based on scale, and they are as follows: brain (1m), topographic maps and systems (10cm), layers and columns (1cm), local networks (1mm), neurons (100 μ m), synapses (1 μ m) and molecules (1Å). Levels of organization in nature and science have been extensively discussed in the first chapter of this dissertation.

A second notion of levels frequently encountered in studying and explaining cognitive phenomena pertains to *levels of processing*, which refers to the stage at which a process occurs within a processing sequence: “Levels of processing are physiological, and refer to the location of a process relative to the transducers and muscles” (Churchland & Sejnowski, 1992, p. 18). The closer it is to the periphery, i.e., to the sensory input, the lower the corresponding level. The order of these levels is governed by a temporal sequence, often connected through cause-and-effect relations. The adoption of levels of processing in biological cognitive systems is the result of considering the mind as an informational processing system (Simon, 1979). The notion of “levels of processing” is commonly discussed in the field of neurobiology, particularly in visual perception and across sensory modalities. In visual perception, low-level processing in the retina transduces light to neural stimuli, intermediate-level processing in the primary visual cortex involves tasks such as contour integration (which boundaries and surfaces belong to specific objects and which are part of the background), and high-level processing in the inferior temporal cortex serves object recognition (a combination of vision and cognition) (Kandell, Koester, Mack, & Siegelbaum, 2021). Here, the term “levels of processing” specifically refers to the stages of neural processing as sensory information travels from sensory receptors to higher-order brain regions responsible for perception and interpretation⁴⁸. As Churchland and Sejnowski (1992, p. 23) put it:

The underlying assumption is that the greater the distance from cells responding to sensory input, the higher is the degree of information processing. Thus the level rank assigned is a function of synaptic distance from the periphery. On this measure, cells in the primary visual area of the neocortex that respond to oriented bars of light are at a higher level than

⁴⁸ Frequently, the term 'stages of processing' is used interchangeably with 'levels of processing.' 'Early stages' correspond to 'low-level processing,' 'intermediate stages' correspond to the 'intermediate level of processing,' and 'late stages' correspond to the 'upper level of processing.' However, using the term 'level' imparts a qualitative distinction to the stages. In the initial stages or levels, the processed representation is still primitive, at a low level of complexity and completeness. As we progress to the next stage or ascend to a higher level, the processing involves finer and more specialized features, resulting in a more integrated representation.

cells in the lateral geniculate nucleus (LGN), which in turn are at a higher level than retinal ganglion cells.

Consequently, this hierarchy exists in all the senses that constitute the perceptual system, commencing with the sensory receptors that detect and transduce sensory stimuli at the lower level, for example, in the eyes for vision, in the ears for audition, and in the skin for tactile sensation. These receptors are the first to interact with sensory information, which is why they are positioned at the lowest level of the hierarchy. At this level, the encoded information is relatively basic compared to the subsequent stages. The processing continues in the primary sensory areas, such as the primary visual cortex in the occipital lobe for vision and the primary auditory cortex in the temporal lobe for audition. These primary areas are responsible for basic feature extraction and the initial encoding of sensory data. Beyond the primary sensory areas, there are secondary sensory areas that receive input from the primary ones. These regions are responsible for the more complex processing of sensory information, such as the integration of different sensory modalities and the extraction of meaningful patterns. For example, in vision, the ventral and dorsal streams in the visual cortex are associated with object recognition and spatial processing, respectively (Kaplan, 2008). In higher-order brain regions, such as the superior colliculus and superior temporal sulcus, where sensory information from various sources is combined to create a coherent perceptual experience, multisensory integration comes into play. Lastly, at the highest levels of processing, association areas in the brain, including the prefrontal cortex and parietal cortex, play a crucial role in cognitive processes related to sensory perception. These areas are involved in attention, memory, decision-making, and the integration of sensory information with other cognitive functions⁴⁹. The hierarchical organization of sensory processing allows for the progressive extraction of more complex and abstract features from sensory input. As information ascends through the hierarchy, it becomes increasingly integrated and cognitively enriched, ultimately leading to the formation of perceptions and conscious experiences⁵⁰. This is a property that concerns mainly information processing systems. In these systems, level ranking depends not only

⁴⁹ For a more extensive description of the mentioned processing stages, consider Kandell, Eric R., John D. Koester, Sarah H. Mack, & Steven Siegelbaum (2021).

⁵⁰ It's important to note though, that the exact organization of sensory processing hierarchies can vary depending on the sensory modality (e.g., vision, hearing, touch) and individual differences.

on the place that they are in a processing sequence but also on the degree of complexity and integration of the processed information.

One of the fundamental advantages of cognitive science, owing to its multidisciplinary nature, is its ability to examine and analyze a cognitive phenomenon from various perspectives and combine the results of these analyses into a comprehensive explanatory theory. These different perspectives of study are referred to as *levels of analysis*, which is the third notion of levels we will explore here. Analysis, in general terms, is a preliminary stage in explaining a phenomenon and it is commonly associated with breaking something down into its components (analysis by decomposition) and the examination of its various aspects. Although this level's notion has a long history, as we delve into contemporary efforts to study a phenomenon, we observe an expanded scope of the concept and the emergence of various analytical methods, both in philosophy, where analysis was first developed and flourished and in various sciences distinguishing different levels corresponding to various methods of studying the same problem (Beaney, 2021).

When defining levels of analysis, it is important to understand that they relate to the focus of a research methodology, which varies significantly across different scientific disciplines. For instance, in the social sciences, levels of analysis refer to the various ways in which society and culture can be studied. These levels range from micro to macro, encompassing everything from the detailed examination of face-to-face interactions to the broad investigation of large-scale historical processes that shape entire civilizations. Typically, sociologists break down the study of society into four separate levels of analysis: micro, meso, macro, and global (Little, 2023). At the lower micro-level, interpersonal interactions are studied. At the intermediate meso-level, the object of study is relationships, dynamics, and interactions within a group. The macro-level concerns much larger groups, such as the population of a city, and the highest global level involves studying phenomena on a worldwide scale, such as the study of social dynamics developing on Facebook or, more broadly, on the Internet. In cognitive neuroscience, levels of analysis focus on different ways of studying a phenomenon, such as the division into three levels mentioned by Churchland and Sejnowski (1992, p. 18): "Levels of analysis are conceptual, and refer to different kinds of questions asked about how the brain performs a task: into what subtasks does the brain divide the tasks, what processing steps execute a subtask, and what physical structures carry out the steps." Other distinctions into levels apply to specific areas of cognitive science, such as the study of language comprehension, where two levels of analysis are distinguished: syntactic, which pertains to the

grammatical structure of a sentence, and semantic, which pertains to its meaning (Eysenck & Keane, 2005).

Certainly, levels of analysis provide a conceptual framework for investigating cognitive processes. They represent a way to structure the study of cognitive phenomena by considering them from multiple angles, each with a different focus and degree of detail. At each level of analysis, researchers ask different kinds of questions to gain insights into how the brain performs a specific task. These questions help break down the cognitive process into manageable components for investigation. The most common division into three levels involves an initial subtask decomposition where a complex cognitive task is broken down into subtasks; for example, if studying language comprehension, one might examine the subtasks of phonological processing, syntactic analysis, and semantic integration. A second level focuses on identifying the processing steps that execute each subtask. For language comprehension, this could include processes like word recognition, parsing, and semantic integration (Eysenck & Keane, 2005). The final level explores the physical structures within the brain that carry out the processing steps. Researchers seek to understand which brain regions and neural circuits are involved in each processing step. These levels are hierarchical, with each level building upon the previous one. They are also complementary because insights gained at one level often inform and enrich the understanding of the other levels.

Today, when we mention the three levels of analysis in cognitive science, we usually refer to the trilevel approach of Marr (1982), which is based on the conception of levels in computer science and analyses a phenomenon (or problem) into the following three levels⁵¹:

Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?

Hardware implementation: How can the representation and algorithm be realized physically? (Marr, 1982, p. 25).

According to Bechtel (2008), these three levels answer two kinds of questions: the first is a what-question and concerns what the cognitive system is doing; namely, it concerns the goal of the computation and focuses on the mechanism in context: "The visual system receives information about wavelength and intensity of light reaching the retina and must, for example, determine the identity or trajectory of the object confronting the organism." (p. 26). The second is a how-question, namely, how it does it: "The two other perspectives Marr proposed focus on the inside of the mechanism and how its components enable the mechanism to perform the task identified in the computational theory." (p. 26)

⁵¹ It has been frequently argued that Marr's use of the term "computational level" is not entirely appropriate, as it suggests a focus on computation rather than the content of a mental state. Among others, Sterelny (1990) has proposed an alternative terminology for the three levels of a psychological theory, which directly correspond to Marr's three levels: "The top level is the level at which we specify the function a psychological device performs. It is the level at which we say what a system does. One down is the level at which we specify the information processing means by which the system performs its function or functions. At the base level we specify the physical realization of that information processing device" (p. 43). In this account, Sterelny refers to the top level as the 'ecological level.' The cognitive capacities specified at this level are not committed to any particular engineering approach, meaning the ecological level remains neutral on mechanism. Importantly, Sterelny notes that "ecological descriptions can be applied to all units in a homuncular decomposition of the mind" (p. 45), suggesting this top level can be addressed at various levels of biological organization, not just at the level of the entire agent. It's worth mentioning Sterelny's critique of various terms used to address the ecological level. He rejects Newell's 'knowledge level' and Dennett's 'intentional' level, but particularly disagrees with Marr's 'computational' level, finding it "odd, given his own agreement that the ecological level is neutral on the means by which information is processed" (p. 46). Below this level, Sterelny identifies two further levels: the computational level, which concerns theories of the cognitive mechanisms that perform specific tasks, and the level of physical implementation, which deals with how a computational process is physically realized.

Marr's trilevel approach was initially developed with the primary aim of understanding the cognitive operation of visual perception, particularly the process of visual object recognition. Since then, it has been applied to various cognitive operations and processes within cognitive science, including language processing, memory and learning, attention and perception, decision-making, and motor control. Because the philosophical literature related to Marr's levels is almost exclusively dominated by examples related to visual perception, let us also examine an example of Marr's trilevel approach related to a different cognitive operation so that we can scrutinize Marr's framework from a different perspective. Let us then explore Marr's levels in the case of motor control, where the point is to understand why we move the way that we do.

In a study aiming to review how the lesion approach and theories of motor control can mutually inform each other (Shadmehr & Krakauer, 2008), Marr's levels analysis was applied to combine the computational, algorithmic, and implementational aspects of this inquiry. The idea of the study was to utilize computational models to identify distinct motor control processes and to map them onto specific neurological disorders, such as lesions in the corticospinal tract, cerebellum, basal ganglia, and parietal cortex. Beginning with the computational level, by using computational models, distinct motor control processes were identified to understand how motor control works. Such processes, linked to specific brain areas, are, among others, system identification, where the cerebellum constructs internal models that predict sensory outcomes of motor commands and corrects them through internal feedback. The parietal cortex, in contrast, integrates predicted proprioceptive and visual outcomes with sensory feedback to form a belief about how the commands affect the states of the body and the environment. The basal ganglia, on the other hand, play a role in optimal control that involves learning costs associated with sensory states and estimating the "cost-to-go" during the execution of a motor task. Finally, the primary and premotor cortices are related to implementing the optimal control policy.

The above are the computational problems that should be solved. Here comes the algorithmic level into play. For instance, in the case of the cerebellum's function of system identification, it was suggested that this computational problem can be solved by building internal models that can predict the sensory outcome of motor commands. Additionally, in the case of the parietal cortex, where the computation concerns state estimation, the problem can be solved by integrating the predicted proprioceptive and visual outcomes with sensory feedback to form an accurate belief about how the commands affect the states of the body and the environment. The last, physical implementational level is involved through the utilization of neural imaging techniques, such as

fMRI, to identify the regions of the brain that correspond to specific motor control processes. Hence, through this analysis, it is evident that Marr's approach is extremely useful in synthesizing existing knowledge and providing insights for further research.

Zenon Pylyshyn (1984), in an attempt to answer a particularly challenging question concerning the precise application of the term 'cognition,' arrived at a hierarchy of three levels, which he labels *levels of description*. Interestingly, these levels bear a notable resemblance to Marr's three levels in the context of the information processing system's analysis, and this is the main reason for referring to them in the same context of levels of analysis.

His reasoning goes as follows: Firstly, he assumes the use of specific descriptions that can capture the generalizations and categories involved in explaining cognitive phenomena. Within this framework, he distinguishes a type of description which he calls 'cognitive description,' distinct from physiological, behavioral, or phenomenological descriptions. It pertains to interpreting the behavior of people in a way that will attribute regularity to their behavior. This description connects beliefs, desires, and human actions to the belief-desire-intention/action schema, a well-known generalization among cognitive psychologists that underlies intelligent behavior. However, beyond this interpretation of folk psychology, human intelligent behavior is also the product of brain function, so how do psychological or cognitive generalizations relate to the underlying processes of the brain? At this point, the analogy of the mind as a computational machine that processes symbols that stand for things (or represent things) outside the brain, things that we believe, desire, intend, set as goals, and the like, comes into play. Building on this analogy, Pylyshyn posits the adoption of three parallel levels of description to elucidate cognitive generalizations used in explaining intelligent behavior: "the biological (or physical) level, the symbolic (or syntactic or sometimes the functional) level, and the semantic (or intentional) level" (Pylyshyn, 1984, p. 259). He further emphasizes that "although the vocabularies of these three levels are distinct, the levels interact in well-defined, principled ways; regularities at one level occasionally show through and modulate the regularities of other levels" (ibid). The introduction of these three levels complements the view of cognition as computation and the mind as a symbol-processing system, which is one of the fundamental assumptions that have shaped the explanations provided by cognitive science.

Some researchers (Boone & Piccinini, 2016; Lotem & Shagrir, 2018; 2021) have pointed out the importance of a different distinction of levels that cognitive science research targets, which bears some resemblance to levels of analysis and levels of description. The levels they mention are two:

one concerns anything related to the content of cognitive phenomena, such as the property of the mind to perform computations and represent internally elements of the external world, and the other concerns anything related to the implementation of cognitive phenomena. This distinction is useful, as it places at one level, often referred to as “functional” (Boone & Piccinini, 2016), all these elements that differentiate cognitive from other biological explanations, while at the other level, “structural” respectively, remain all which are common to biological cognitive and purely biological phenomena. Furthermore, it enabled the study of the computational properties of an information processing system independently of its physical implementation. This is related to the principle of autonomy between functional and structural explanations, the validity of which has been the subject of significant disagreement within the realm of the philosophy of cognitive science.

The final type of levels that I'll mention, originating from the field of philosophy and closely tied to levels of analysis, concerns *levels of realization*. “Realization,” a concept that emerged in the philosophy of mind in 1960 through Putnam's classic paper entitled “Minds and machines” (Putnam, 1960), is a rather technical term used in metaphysics, philosophy of mind, and philosophy of science to indicate a relationship believed to exist between higher-level properties and lower-level properties (Baysan, 2015). It relates different kinds, for example, a psychological to a physical kind, like pain and its associated brain state, but it should be kept in mind that both are properties of the same thing. In the pain example, the brain state is the realizer, and pain is the realized property. According to the realization principle, mental properties are dependent on physical properties, functional and computational properties are dependent on underlying first-order properties that fulfill specific causal or functional roles, and dispositional properties find their basis in categorical properties, and so on. The term *realization* is typically associated with the principle of multiple realizability, and in the field of philosophy of mind, this means that every mental property can be actualized by different physical properties, which are substrates that serve the same role as that of the brain. In any case, this umbrella term refers to a dependency relationship between higher-level properties or states and lower-level properties or states in the following sense: “The instantiation of a higher-level property or a state depends on, and is necessitated by, the instantiation of its lower-level realizer (or realizers)” (Baysan, 2015, p. 248).

But how does the above relate to the explanations provided within cognitive science? Levels of realization, used mainly in the philosophical literature and terminology, in contrast to the three previous notions of levels, are not encountered as frequently in scientific texts related to cognitive

phenomena⁵². Realization levels are related to cognitive science by expressing the underlying assumption connected to the use of the term *neural correlate* (Wilson, 2004). Typically, realization is implied through the silent metaphysical assumptions of cognitive science concerning how the mind is related to the brain. The most fundamental such assumption, which is also related to the broader idea of science, is physicalism, which in its minimal form is the claim that “every property of a thing is either a physical property or is determined by its physical properties” and that “there is nothing in the world that is not a physical thing” (Kim, 1998, p. 12). This position represents the combined acceptance of the following three principles:

[Mind-body supervenience] The mental supervenes on the physical in that any two things (objects, events, organisms, persons, etc.) exactly alike in all physical properties cannot differ in respect of mental properties. (Kim, 1998, p. 10),

[The anti-Cartesian principle] Nothing can have a mental property without having some physical property and hence without being a physical thing. (Kim, 1998, p. 11), and

[Mind-body dependence] What mental properties a given thing has depends on, and is determined by, what physical properties it has. (Kim, 1998, p. 11)

As a result of embracing physicalism and the multiple realizability of mental properties, the scientific study of cognitive phenomena is focused on two axes: First, the mental content processed in cognitive phenomena is examined under the metaphysical umbrella of functionalism.

Functionalism is a consequence of the multiple realizability of mental properties in physical systems and asserts that mental properties are not bound by how a physical or biological system implements them. It proposes that, due to this, mental kinds can be examined abstracted from the physical substrate that realizes them and should be considered as causal functional kinds, which are internal states caused by certain sensory inputs and can result in specific behavioral outputs (Kim, 1998). Second, the physical transformations of the physical substrate that implement mental kinds mainly concern biological organisms, and relevant research aims to discover the underlying

⁵² Wilson (2004) suggests that there is a conceptual gap between scientific and philosophical usage of the concept of realization. He notes that although cognitive scientists use the term 'realization' to describe the physical mechanisms underlying psychological functions, they do not use the term in the same way that philosophers do: "The short answer to the question “How do cognitive scientists conceptualize realization?” is “They don’t.” Or, to put it more accurately: they do, but almost exclusively when they are attempting to sketch a broader location for their particular views or to interact directly with their philosophical interlocutors.” (p. 987)

neurobiological mechanisms. This approach aligns perfectly with the physicalist commitment to scientific research. However, what remains to be answered here is how the first category, which pertains to mental content and processes, relates to the second, which deals with physical entities and processes. What is their relationship? Do they cause one another? Are they the same or different? Here, the realization relation answers that they are properties of one and the same thing. While this discussion about realization could lead to numerous metaphysical implications, such considerations are beyond the scope of this dissertation. Therefore, I will adopt a metaphysically neutral stance towards the realization relation. Primarily using the term “neural correlate” when referring to the physical basis of a mental state allows us to explore the relationship between mental and neural states without becoming entangled in complex metaphysical debates⁵³.

The notions of levels discussed above, along with the various organizational levels applied in the biological sciences (as extensively described in the first chapter of this dissertation), represent the most common cases in which hierarchies are encountered in cognitive science. These levels are reflected in the hypotheses, theories, explanations, and research methodologies employed to investigate cognitive phenomena. But how do these level notions relate to one another? And how do they intertwine to form unified explanations in cognitive science? To explore these questions in greater depth, we will examine a specific case study: the process of seeing and recognizing an apple. This example from visual perception will allow us to illustrate how different uses of levels in an explanatory attempt come together to provide a comprehensive understanding of a cognitive phenomenon.

⁵³ While the present discussion about realization involves various metaphysical positions regarding the relationship between mental and physical properties, recent work in the philosophy of neuroscience (Chirimuuta, 2024) suggests that a more neutral terminology might be preferable. Interestingly, Chirimuuta argues for an approach she terms “haptic realism,” emphasizing that scientific knowledge is actively constructed through interaction with nature, rather than passively received. This perspective cautions against making strong metaphysical claims based on neuroscientific findings. Instead of terms like “realization,” “supervenience,” or “determination,” which carry specific metaphysical implications (for instance, positions such as epiphenomenalism, panpsychism, physicalism, and so on), Chirimuuta’s approach would favor the use of “correlate” when describing the relationship between neural and mental states. This term acknowledges an observable relationship without committing to specific causal or ontological claims. Such a stance recognizes the complexity of brain-mind phenomena, the limitations of our current scientific understanding, and the need for multiple perspectives in neuroscience, as is also argued by Craver (2007). By adopting this more neutral terminology, researchers can focus on empirical findings while remaining open to various theoretical frameworks, thus avoiding premature metaphysical commitments in a field where, as Chirimuuta argues, simplification and schematization are necessary but potentially distorting features of scientific practice.

5.2.2. Visual Perception as a Model for Multilevel Explanations in Cognitive Science

Before exploring the case, a widespread view among those studying cognitive phenomena will be shortly examined: for some, trilevel Marr's analysis corresponds to levels of realization. If this is true, then in cognitive science explanations, we are not dealing with four different notions of levels but essentially with three, namely levels of organization, analysis, and processing. Horgan and Tienson (1996) argue that Marr's levels account shares many common characteristics with realization relation:

The relationship between state types at Marr's middle (algorithm) level and state types at the lowest (implementation) level, and also the relationship between state types at Marr's top (cognitive-transition [computational]) level and state types at his middle level, is the relation that philosophers call realization and cognitive scientists call implementation. (p. 23)

And Craver (2007) explicitly states:

David Marr's (1982) 'levels of analysis' [...] are perhaps the best-known example of levels of realization. Computational processes, such as addition, are realized by algorithmic transformations, and these, in turn, are realized by an organized collection of hardware components. These are three different properties of one and the same system. The system as a whole is at once an adder, a manipulator of symbols, and an organized set of electrical circuits. (p. 218)

The correlation between Marr's levels of analysis and levels of realization, which some consider identical, is not surprising at all. Firstly, both these notions of levels concern one and the same entity. Marr's three levels of analysis pertain to different ways of studying a single entity, e.g., a cognitive system. Similarly, levels of realization concern the relation between different properties of a single entity and the dependencies that concern these properties. The assertion that Marr's levels essentially constitute levels of realization implies that the properties attributed to each higher level are realized at the lower levels. Thus, the ability of a system to exhibit certain behavior, which is ascribed to the computational level, according to Marr, depends on/is realized in its ability to perform certain informational operations, and their execution is realized in the physical substrate of the system. According to others (Polger, 2004), while there is a significant correlation, there is no complete identification, and this specific issue is highly debatable regarding how each philosopher interprets the realization relation.

In general, and within the objectives of this section, which aims to identify the various ways in which different notions of levels enter into cognitive explanations, we can adopt Craver's perspective, which attributes realization relations to Marr's three levels. It should also be kept in mind that these levels' notions have different starting points. Levels of analysis pertain to different ways of studying a system, while levels of realization pertain to the system's nature and how it is manifested through its properties. Hence, the first type has a methodological and epistemic character, and the second a metaphysical one. At the end of the day, however, what interests us here is to examine how the content of a cognitive state correlates with its physical implementation by the brain within the framework of an integrated multilevel explanation. The use of the trilevel analysis, given that it captures the essential role of mental content, serves its purpose well. Whether we are referring to levels of realization or levels of analysis, the content pertains to the higher level. It is the mental property that is realized in processes embodied in a specific physical substrate, or in Marr's terms, it is the computational goal executed by an algorithm physically implemented in a neurobiological structure. Hence, if we aim to apply the earlier discussion to illustrate a cognitive operation, specifically the visual perception and recognition of an object (e.g., an apple) while adhering to the underlying assumption of the computational theory of mind upon which Marr's view and the realization concept is based, we will move on to the following observations.

Let us first examine how levels of realization are involved in this specific exemplar explanation and to which Marr's levels of analysis correspond. At the higher level concerning the behavior of the studied (visual) system, the mental property under consideration is that the system perceives an apple. This, at a second, lower level of realization, is based on the property of the visual system to act as a "computer" that receives as input a proximal stimulus, which is the reflection of light from the visual scene onto the eye. Additionally, it relies on its ability to encode the activation pattern of the proximal stimulus into information and to process it within the brain. At a third level of realization, the before-mentioned properties are realized in the property of the visual system to consist of a transparent sphere containing a lens and a photosensitive surface, along with networks of neurons, with specific arrangements, organized in a particular way in the brain to carry out specific processes. All three descriptions relate to properties of the same entity, namely, the visual system: first, the visual system perceives (the apple), which is a form of computation; second, the visual system processes information; and third, the visual system has a specific structure that enables it to perform these functions.

In this manner, we have already divided the explanation into three levels of realization or analysis. We will further explore these levels by utilizing Marr's (1982) trilevel approach to gain a more thorough understanding of the operations involved in the visual perception of an apple. At the upper computational level, which pertains to what the visual system does in relation to its environment, the answer is that it perceives an apple. The underlying algorithmic and representational level answers the questions of how it does this and what representation is used for the input-to-output transformation. The answer is that it perceives the apple visually through several operations. More specifically, it encodes the image -light reflected from the apple and its surroundings- into activation patterns where the image spatial structure is preserved (in visual field maps), initially in the early stage of visual processing (regarding the retinal circuitry and subsequently the primary visual cortex, viz, V1). In this stage, basic features of visual stimuli, such as orientation, spatial frequency, and color, are analyzed (low-level visual processing). The extracted information then proceeds to the next stage of processing, residing at the extrastriate cortex (V2, V3, V4, V5, and MT), where basic features like contour, color, and motion are assembled into a unified percept, such as the shape of the apple against a background (intermediate level visual processing). Here, too, there is a degree of preservation of the image's spatial structure (Wandell & Winawer, 2011). At each processing stage, the characteristics of the image are grouped more and more until they reach higher levels of processing. At this stage, the related brain areas include the Inferior Temporal Cortex (IT), the Lateral Occipital Complex (LOC), and the Fusiform Face Area (FFA), where more complex and integrated stimuli, such as the shape of a face, a hand, or a sphere are processed and with the involvement of cognitive processes like memory and attention, the apple is recognized and identified. The physical implementational level addresses where and by what means the described representations and algorithms are executed. In other words, it explores the physical characteristics of the involved brain regions. These pertain to (1) the connectivity patterns that enable the gathering of activation patterns from different areas of the visual system, (2) the involved local networks responsible for the distribution of inputs and outputs and feedback loops, (3) features regarding certain cells as is the responsiveness of cell receptive fields to various types of stimuli such as motion, orientation, color, or more complex shapes. All the above physical structures are responsible for executing the algorithm that accomplishes the computation that is described and analyzed at the upper computational level. This is how the story unfolds across Marr's levels of analysis.

Let's now extend this explanation in a more detailed version to explore how Marr's levels, levels (or stages) of processing, and levels of organization could be intertwined in a multifaceted explanatory description. For the purposes of the example, I will focus the explanation on three levels of organization which are, from the lower to the higher, the cellular, the level of local networks, and the higher level of brain regions. At every level of organization, different theories, different research questions, and different methodologies are recruited. The same stands for levels of analysis, too, insomuch as they address different aspects of the same mental phenomenon that is to be explained, for example, what-questions are answered by behavioral or psychophysical studies, whereas how-questions are answered by neuroimaging or cell-recording studies, among other methodologies. With this in mind, the explanation we will explore next is an attempt to include and express all these different aspects used to contemplate a phenomenon and construct a unified, multifaceted, and inclusive explanation.

Let's focus on the lower level of organization, the cellular, and begin with the lower level (early stage) of processing. At this early processing stage, the computational goal is to detect and represent the basic visual features of the apple, such as its edges, colors, and overall shape. At the algorithmic level, to solve this computational goal, the visual system in early processing visual areas collects information from receptive fields that respond to specific features of the apple. Initially, light reflected from the apple is captured and converted into information encoded in electrical signals. Subsequently, visual information is collected and transmitted to the brain, where initial processing and filtering of visual input takes place, and basic algorithms are executed, resulting in edge detection and basic feature extraction. At the physical implementational level, the protagonists of these processes are photoreceptors (rods and cones) for light conversion, retinal ganglion cells, and LGN neurons for visual information transmission, and complex, hypercomplex, and other types of cells located in specific areas in V1 for basic feature extraction. They can execute all these functions due to their physical features and the configuration of their receptive fields.

Ascending levels of organization, let's say that we attempt to explain the same phenomenon at the local network level; here, the computational goal is to enhance edge detection, contour continuity, and feature extraction within a local visual field focused on the apple. The algorithm that implements this computational goal utilizes lateral connections to facilitate contour integration, inhibitory mechanisms to enhance edge sharpening, and feedback loops to contribute to local feature extraction related to the apple. More specifically, these mechanisms are implemented in (1) local circuits comprised of simple and complex cells, with the first having receptive fields

responsive to basic features and the second to more complex patterns and contours, (2) orientation columns in V2 where neurons within a column respond to a specific orientation but vary in their selectivity to other features, (3) hypercolumns in V2 which consist of clusters of neurons responding to different visual features, such as color, orientation, and spatial frequency, and enhance the representation of oriented features and contribute to the continuity of contours, (4) blobs, interblobs and other structures with different roles.

If we proceed with this explanatory description at a third, higher level of organization concerning brain regions, we will detect the computational goal, which is to transform raw visual input related to the apple into a coherent and interpretable representation, progressing from basic features to more complex visual information. The algorithmic level concerns hierarchical processing specific to the apple with feedforward and feedback connections contributing to the refinement of visual information, and at the implementational level, the involved areas are the retina, which contains photoreceptors and ganglion cells, the optic nerve, which transmits visual information from the retina to the brain, the Lateral Geniculate Nucleus (LGN), which is a relay station that receives input from the optic nerve and projects visual information to the primary visual cortex, V1 which contributes to the analysis of basic visual features such as edges, orientations, and colors, V2 which is involved in more complex feature processing and extends the analysis of basic features initiated in V1, and other structures such as superior colliculus, thalamus, e.ct.

A similar analysis can be made for the intermediate level (stage) of visual processing: Here, the computational goal is the hierarchical processing of visual information by further integrating features to represent motion, depth, and form and preparing representations for complex object recognition. At the lower neuronal (cellular) level of organization, this is accomplished via algorithms that exploit lateral connections, contour integration mechanisms, and feature conjunctions to extract more complex shapes and contours related to the apple. These processes are implemented by many types of cells, such as complex cells in V2 and V4, which integrate the responses of simple cells, end-stopped cells in V2, and hypercomplex cells in V2 and V4, for line endings and corners detection, interblob and blob cells, for color and form processing, and other types of cells. At the local network level of organization, the algorithm exploits long-range lateral connections (e.g., connections that link neurons similar in function but representing distant locations in the visual field) to facilitate contour integration specific to the apple, grouping mechanisms that aid in feature conjunction and convergence/divergence of information that refines the representation of more complex shapes. These are implemented in local networks that

connect neurons at several areas, for example, the ventral cortical pathway, which connects V1, V2, and V4 and terminates at IT for form and color perception, and the dorsal pathway that terminates in MT and MST and serves motion and depth perception. Lastly, and within the brain region level of organization, all the intermediate-level processes take place at many cortical areas, including V1, V2, V3, V4, and Middle Temporal area (MT). They are involved in integrating local cues to construct contours and surfaces and segregating foreground from background.

If we continue with the higher level (or stage) of processing, the computational goal is the formation of high-level object representations, the integration of information from various visual pathways, and the support of recognition, categorization, and interpretation of complex visual stimuli that convey the representation of an apple. The algorithm for this is mainly based on mechanisms for the formation of invariant representations by utilizing object constancy mechanisms. For high-level processing tasks, almost the whole brain is engaged with several areas and processes that contribute to the computational task of visually perceiving and recognizing the apple. The Inferior Temporal (IT) cortex, with neurons that respond selectively to complex object features, is the primary center for object recognition; the Lateral Occipital Complex (LOC) integrates information about object shape, orientation, and other features, while the Fusiform Face Area (FFA), apart from primarily processing faces is also involved in the high-level processing of other complex visual stimuli, including objects. Other areas concerning cognitive functions, such as short-term and long-term memory, attention, and motor control of saccadic movements, are also engaged⁵⁴ in this stage of processing.

I appreciate that the general idea has been conveyed by the fairly detailed example I presented: different aspects of the explanandum, which pertains to what is perceived and how, different stages corresponding to the increased complexity of the image represented, and different levels of organization, all of them deriving from different research methodologies, all together contribute to answer in a complementary manner, how someone perceives and recognizes an object. Certainly, a comprehensive explanation necessarily involves various levels of organization. It encompasses lower levels and focuses on the microscopic subcellular structures, such as the internal architecture of photoreceptors, or examines the role of rods and cones and ganglion cells in the retina.

⁵⁴ For a detailed description of the mechanisms and areas involved in the three levels of visual processing, consider chapters 21, 22, 23, 24, and 25 in *Principles of Neural Science* (Kandell, Koester, Mack, & Siegelbaum, 2021).

Ascending levels of organization, an explanation involves larger structures like blobs and columns, where cells are organized based on their functional characteristics and connectivity, as is the case with ocular dominance, orientation, and cortical columns. As we further ascend levels, we'll find even larger brain structures specialized in perceiving specific types of entities, such as the Fusiform Face Area (FFA), Lateral Occipital Complex (LOC), and Inferior Temporal (IT) cortex, closely associated with object processing. In each processing stage, all levels of organization are necessary for understanding the relevant mechanisms, and the explanation cannot be confined to just one. In other words, one cannot elucidate the process of vision solely in terms of neurons or brain regions. These distinct levels of organization are essential for understanding what and where questions and moreover, they jointly define the algorithm and the implementation of the computational goal. Importantly, since the explained processes involve cognitive phenomena with mental content, all the aforementioned levels must be correlated with this content. As illustrated in the example, this correlation is achieved by including in the explanation the higher computational (Marr's) level. Finally, a thorough explanation necessarily extends across all levels of processing because the extraction of information in the initial processing stages is coarse, and the generation of the integrated representation corresponding to the perception of the visual scene is completed only at the final stages.

Returning to the distinction between content and vehicle introduced earlier in this chapter, it is not difficult to observe how it fits into the example of explaining the visual perception of an apple. The content pertains to what the subject sees, that is, an apple. Internally, it is the representation of the apple gradually constructed through various processes in the visual system corresponding to specific informational content. According to the case study presented, this representation, at the beginning (low-level) of processing, is fragmented into smaller, simpler, elementary contents. In other words, the informational content initially processed by the system involves dots, spots, lines, edges, and similar informational entities. However, as we progress to higher processing levels, this information is organized into more complex patterns, such as contours and specific forms, which are contextually recognized at the upper level of processing - along with the involvement of other cognitive processes - and identified as the specific representational content, which is the image of an apple.

Mental content corresponds to the question 'what'. What do you see? What do you taste? What do you remember? Since the presented example pertains to visual perception, the answer and the corresponding explanation relate to the behavior [*sees an apple*]. Other types of mental behavior

with the same content could be the object of a similar discussion. For instance, the behavior [*remembers an apple*] involves operations and structures of the memory system. Similarly, the behavior [*recognizes tactually an apple*] would encompass operations of tactile perception. In all three cases, the content is the apple, constructed by processing specific characteristics of the apple. For visual perception, features like shape, color, and position constitute its representation, specializing the content related to visual perception and constituting its form, pertaining to its representational vehicle. If we explain the memory of an apple, additional features of the content, such as taste and crispness, should be included to structure another format of the apple's representational vehicle. Similarly, in tactile perception, the representational content involves processing of the shape, texture, temperature, weight of the apple, etc. In all these cases, the mind executes specific mental operations that involve the transformation of symbolic representations or mental symbols. Thus, the visual system manipulates mental symbols that stand for properties of things in the outer world, such as the color, shape, and position of the visual object. Through subsequent processing stages, it combines these symbols, or contents, into a more complete representation of the seen object. The mentioned contents constitute computational goals that are determined accordingly for each level of processing and organization, as illustrated in the case study presented here. Correspondingly, the processes and related neurobiological structures described partially at the algorithmic and at the physical implementational levels are aspects of the vehicle.

With these clarifications, we are ready to proceed to the next section, which will explore how cognitive explanations fit into the new mechanistic framework. The focus will be on the mechanistic levels.

5.3. RECONCILING COGNITIVE SCIENCE EXPLANATIONS WITH MECHANISTIC LEVELS

In the field of cognitive science, researchers strive to unravel the mysteries of the mind, exploring different levels of organization, analysis, and processing. Trying to apply the new-mechanistic framework, which focuses on mechanisms, various levels, and the integration of different research approaches to cognitive phenomena, a key question arises: How do mechanistic levels align with the levels commonly encountered in cognitive science? Answering this question will likely provide

us with an insight into whether the constitutive mechanistic framework for explanations can efficiently accommodate the multilevel character of cognitive explanations.

Up to this point, we have examined two fundamental aspects that differentiate the scientific exploration of cognitive science from exploring other biological phenomena. The first aspect regards what we have seen in the first section of the chapter: We observed that cognition concerns systems where, in addition to physical processing, informational processing takes place, with the particular meaning that information has in cognitive systems, viz, systems exhibiting intentionality. This parameter was explored through the distinction of content versus vehicle. The second aspect pertains to the interdisciplinary and multilevel nature of cognitive science explanations, extensively examined in the second section by scrutinizing different notions of levels used in the study and explanation of cognitive phenomena. This multilevel character of cognitive science explanations was further developed with the aid of an example. Now, the critical question arises: How do these two aspects relate to what the new-mechanistic explanatory framework advocates? Regarding the first aspect, a physical state that possesses intentionality seems to pose a challenge to the applicability of the new-mechanistic framework to cognitive phenomena. The reason is that a comprehensive mechanistic explanation should include the role of processing informational content and generating internal representations. Therefore, the question is, if this can be achieved and how. On the other hand, the second aspect we explored, i.e., the multilevel character of cognitive phenomena and explanations, appears to be relevant to the framework of constitutive mechanistic explanations because these explanations, too, explore a phenomenon in a hierarchical manner that reflects the various methodologies encountered in each level. On first reading, then, such an explanatory framework has the potential to align with the multilevel character of cognitive phenomena. However, there are strong reasons to maintain that this is not the case. And this has to do with the different notions of levels we examined in the previous sections, whose understanding significantly contributes to identifying the challenges in applying mechanistic levels in cognitive science, at least in the form presented by Craver (2007).

5.3.1. Revisiting Levels of Mechanisms: Strengths and Limitations

Let's take a closer look at the mechanistic levels, often referred to as levels of mechanisms, as outlined by the new-mechanists. The aim is to explore their relation with the levels encountered in the theories and methodologies of cognitive science. In the literature associated with the new-mechanistic movement, various attempts to define levels of organization for explaining biological

phenomena are extensively discussed and criticized. The reason is that these attempts fall short when used to explain the mechanisms generating and underlying biological phenomena (Povich & Craver, 2017; Bechtel, 2008; Craver, 2007) because they primarily focus on more generalized properties, such as size, intensity, and duration of the entities or processes they attempt to describe. In contrast, mechanistic levels of organization exhibit a much more specific and applied nature, and therefore, they are tailored to the explanation per se. Among the efforts to define levels for mechanistic explanations, Craver's (2007) account stands out as the most comprehensive. Craver provides an extensive presentation of what is meant by mechanistic levels, introducing a critical element that sets his approach apart from many other level accounts prevalent in the field. This crucial element lies in Craver's definition of mechanistic levels as *levels of mechanisms* (Craver C. F., 2007, p. 163). As further articulated by Povich & Craver (2017):

In constitutive mechanistic explanations... one looks to a lower level, within the phenomenon, to reveal its internal causal structure. In many such systems, one can explain the behavior of the whole in terms of the organized behaviors of its parts, and one can explain the behaviors of the parts in terms of the organized behaviors of their parts. Levels, on this view, are simply embedded mechanistic explanations (p. 186)

Emphasis should be placed on the last sentence, which states that “Levels, ..., are simply embedded mechanistic explanations”. What does this statement entail? When examining attempts to define levels based on criteria like size, intensity, interactivity between entities, parthood, etc., we observe a common thread: all these definitions presuppose an initially established hierarchical system of levels where operations are described within a pre-existing framework. To illustrate, consider the attempt to establish three levels of biological organization in humans: organs, tissues, and cells. Within this framework, the function of the lungs can be explained as an organ interacting with other organs (e.g., the heart), as a tissue composed of alveoli, and as a set of cells arranged in the alveolar epithelium to facilitate the filtration of incoming air and the absorption of oxygen. In this context, the heart is positioned at the same level of organization as the lungs (organ level), but also alongside the brain, liver, and spleen, as they are all organs. Similarly, at a lower level are the heart's muscle cells, alveolar cells, neurons, and every other cell type in the body, categorized at this level due to their cellular nature. In this specific example, the criteria for classifying a structure at a level beyond morphological characteristics include size and part-whole relations. It's crucial to note that the entire hierarchy of levels was presupposed before developing the paradigm described above.

However, if we follow Craver's (2007) and Povich and Craver's (Povich & Craver, 2017) reasoning concerning mechanistic levels, we understand that these levels are not predetermined but rather emerge from the process of mechanistic explanation. This is evident in the terminology *levels of mechanisms* embraced by the new-mechanists. Throughout the explanatory process, underlying mechanisms are revealed, and each underlying mechanism that explains a phenomenon is automatically placed at a lower mechanistic level -in other words, it is a lower-level mechanism- always in relation to the phenomenon it explains and which may be part of another, higher-level mechanism. In the context of the new-mechanistic perspective on levels of explanation, let's consider the example of the heart and the lung. The cells of the heart are not at the same mechanistic level as the cells of the lung because they contribute to different mechanisms: blood circulation and respiration. To put it another way, the philosophy of positioning an entity or a process at a level differs significantly between predetermined levels of organization and mechanistic levels of organization. While for the former, it is sufficient that something is an organ or a specific type of cell to be placed at the same level as another organ or another type of cell, for the latter, e.g. mechanistic levels, if the entities do not belong to the same mechanism, we cannot, under any circumstances, claim that they belong to the same level. With the iterative application of explanations, or mechanisms, which involve the description of spatially and temporally organized entities and activities, a hierarchy of levels is established, specific to certain biological phenomena. This hierarchy essentially reflects a hierarchy of mechanistic explanations. As highlighted by Bechtel (2008, p. 147), "a working part of a mechanism may itself be a mechanism," and the quest for working parts consistently leads us to progressively smaller-scale mechanisms and working parts.

For this reason, it becomes much more conceivable why mechanistic levels are considered to be local in character. Considering that these levels derive from mechanisms, it becomes difficult to conceive of their existence independently, outside the hierarchy of mechanisms responsible for their generation. This localized nature of mechanistic levels stands as a significant strength of this framework because it focuses on the explanatory relevance of the properties that define the levels, excluding properties unrelated to the explanation. In other words, the local nature of mechanistic levels means that entities sharing the same level need only be components of the same mechanism, participating in the explanation of a phenomenon or the behavior of a system. Other properties, such as size (for component parts) or duration (for operations), may influence how these entities function or interact. Still, it is crucial to emphasize that it is not their size or duration that

determines their placement on the same level. Instead, it is the shared participation and interaction within the same mechanism that establishes their level equivalence.

Another element that makes the idea of mechanistic levels distinct from other levels' conceptions is that they consist of working parts, not just parts. This is linked to the fact that we refer to levels of mechanisms, and since something, in order to be part of a mechanism, must perform some function in that mechanism, every part is a working part. In contrast, alternative accounts of levels of organization may include entities at a certain level without necessitating a functional role. For instance, in a framework defining levels based on the size of an entity, something might be assigned to a specific level merely due to its size, sharing this characteristic with other objects of similar scale, irrespective of its functional contribution.

Despite not being grounded in criteria such as size, intensity, duration, and the like for their establishment, mechanistic levels do, in fact, satisfy these criteria as a natural outcome of their inherent nature as levels of mechanisms. As highlighted by Povich & Craver (Povich & Craver, 2017), this conceptualization of levels:

...satisfies many of our pre-analytic intuitions about the levels of explanation. Things at higher mechanistic levels are typically larger (and necessarily no smaller) than things at lower levels because the latter are parts of the former. Things at lower levels often take less time than things at higher levels because activities at lower levels compose activities at higher levels. And the idea of mechanistic levels captures the common idea of “levels of organization” because higher-level mechanisms are made up of organized parts and activities. (p. 188)

It is essential to grasp that this quality is a secondary feature that causes mechanistic levels to bear a resemblance to commonly defined levels of biological organization; in other words, they serve a role as levels of composition but are not identical. As underscored, they are fundamentally levels of mechanisms, inherently local and tailored to specific systems and the phenomena they manifest. Indeed, mechanistic levels are inherently case-specific.

In any case, mechanistic levels belong to the broader category of levels of organization and the narrower category of levels of composition, governed by part-whole relations, as acknowledged by Craver: “Levels of mechanisms, [...] are a variety of part-whole relation” (Craver C. F., 2007, p. 165). This characteristic, namely, belonging to the category of levels of composition, is responsible for difficulties in applying mechanistic levels in the cognitive domain. I argue that precisely because

mechanistic levels are essentially compositional levels, while compatible with biological explanations, they cannot accommodate cognitive explanations in the sense in which mechanistic levels are commonly understood. As demonstrated in the previous section, there are notions of levels involved in the exploration of cognitive phenomena that are significantly different from levels of composition. Therefore, since mechanistic levels are levels of composition, they cannot capture the roles and specific characteristics encountered in other notions of levels, which are essential in cognitive explanations, at least not without making a serious effort to reconceptualize their role in structuring an explanatory framework that includes the mental aspect. In what follows, I will explore these specific difficulties in more detail.

5.3.2. The Challenge of Incorporating Informational Content in Levels of Mechanisms

Craver (2007) elaborates his account of mechanistic levels by presenting a multilevel explanation of spatial memory. In this example, the ability of a mouse to navigate a Morris Water Maze (upper-level of spatial memory) is explained through the mechanism of spatial map formation. This mechanism is detailed at a lower level (the level of spatial map formation), focusing on the hippocampus's function. At this lower level, Craver incorporates the computational properties of neural systems. Addressing the question of how spatial maps are generated in the hippocampus, he describes the mechanism of Long-Term Potentiation (LTP) in hippocampal synapses, situated at an even lower level. Finally, at the bottom level of his hierarchical model, he describes the molecular mechanisms that explain LTP. This specific case serves as a reference to demonstrate how levels of explanation contribute to a constitutive mechanistic explanation, supporting Craver's assertions on levels of mechanisms: these levels reflect the way that neuroscience formulates explanations spanning multiple levels. To the extent that this account concerns neuroscience as a study of physical (biological) entities and activities organized to explain the behavior of a biological system, it succeeds in its objective. However, Craver's classification of levels, drawn from this specific example, fails to capture the unique, informational and intentional nature of cognitive phenomena. This limitation arises because informational and representational entities are not incorporated at any point within the underlying mechanistic levels.

This issue becomes apparent when we attempt to explain a cognitive phenomenon using a hierarchy of mechanisms. Let's examine this more specifically. At this point, it would be illuminating to revisit the comparison of the two biological cases presented at the outset of the

chapter (section 5.2.1)⁵⁵. The first case involved the function of muscle contraction in executing a movement, while the second focused on the operation of the visual perceptual system, expounded in greater detail in section 5.3.2. In both biological cases, the explanations unfolded within a hierarchy of mechanisms, spanning from a higher level encompassing observed behaviors (such as the bending of the forearm and the visual perception of an object) down to the level of detailing molecular mechanisms (specifically, cross-bridge cycles and phototransduction respectively).

Both of these explanations concern entities interacting with each other to carry out various activities. Focusing on the first explanation related to the motor system operation, the entities involved include muscles, tendons, the brain, spinal cord, motor neurons, bones, motor units, muscle fibers, motor plates, calcium, sarcomeres, actin, myosin, and more. A noticeable aspect is that these entities belong to various levels of organization, thus indicating part-whole relations between them. For instance, muscles, tendons, the brain, spinal cord, motor neurons, and bones are categorized as body parts and organs. Muscles are further organized into motor units, which, in turn, consist of muscle fibers and motor plates. At a lower level, muscle fibers include sarcolemma, myofibrils, and calcium. Myofibrils, in turn, contain sarcomeres, which consist of actin and myosin filaments. This organizational structure reflects a compositional hierarchy across different levels.

Do these levels align with mechanistic levels? To determine this, we could employ Craver's (2007) questions, which he uses to develop his mechanistic levels account: (1) what are the relata, (2) what is the interlevel relation, and (3) by virtue of what are any two items at the same level (pp. 171-172). In response to the first question, the relata are the working components within the flexion movement mechanism. Each entity identified serves a distinct function: the brain generates a pattern of activation, the spinal cord distributes it to relevant nerves, tendons employ muscle power to facilitate bone movement, motor units are recruited to enhance muscle power, motor plates signal muscles to initiate contraction, calcium is released inside muscle fibers to start the contraction mechanism, and so forth. Moving on to the second question regarding the interlevel relation between the entities involved, it is evident that this is a part-whole relation since every lower-level entity partially constitutes some upper-level entity. What about the third question? Any two items on the list are at the same (mechanistic) level when they are working components of the

⁵⁵ The specific descriptions of the mechanisms were brief. For more details on how these explanations unfold within a hierarchy of levels, refer to section 3.3. (for the muscle case) and section 5.3.2. (for the visual perception case).

same mechanism, and entities at the same level do not exhibit a part-whole relation between them. For example, actin and myosin are both components of the sarcomere's contraction mechanism, and neither is a component of the other. This assessment leads us to a conclusion about the description of the musculoskeletal mechanisms that explain forearm flexion: the mentioned mechanisms form a hierarchy, which corresponds to a constitutive mechanistic explanation unfolded in levels, perfectly aligned with levels of mechanisms.

The previous description of the muscle movement mechanism exemplifies a constitutive mechanistic explanation akin to Craver's (2007) spatial memory example, which he uses to illustrate mechanistic levels. It involves a behavior -the flexion of the forearm- explained by the contraction of the involved muscles. Similarly, Craver's example, with the mouse navigating the Morris Water Maze, involves a behavior explained through brain operations. These two cases, although similar in the methodology used, present a significant difference: one concerns a bodily function, and the other a cognitive operation, yet both explanations utilize solely biological entities. Let us further examine how the explanations unfold at lower and more detailed levels of organization. Starting with the upper level, the mouse example involves maze navigation, while the forearm example involves bending. In the mouse, the mechanism resides within the hippocampus; in the forearm, it resides within the muscle. The mechanism of spatial map formation in the mouse is explained through Long-Term Potentiation (LTP) in hippocampal synapses; in the forearm, the mechanism is explained through the activation and contraction of muscle fibers by the coordinated contraction of their sarcomeres. LTP in the mouse is further elucidated by the molecular activation of the NMDA receptor, while sarcomere contraction is explained by the molecular interaction of actin and myosin filaments, generating cross-bridge cycles. In essence, despite the cognitive nature of the explanandum behavior in Craver's example -linked to memory- compared to the bodily nature of the movement example, there is no substantial difference in how the levels concerning these mechanistic descriptions align with mechanistic levels of organization. Both the mouse example, explaining a cognitive capacity, and the muscle example, explaining a motor ability, utilize the same approach to mechanistic levels.

It is now time to examine the second explanation pertaining to visual perception. As has already been observed, the entities involved in the explanation are not all of the same nature. Specifically, two lists of entities can be compiled. One list contains physical entities such as the eye, optic nerve, primary visual cortex, lateral geniculate nucleus, retinal circuitry, ganglion cells, rods, cones, and photopigments. The second list consists of entities such as images, objects, movement, shape,

boundaries, orientation, contrast, contour, lines, edges, and line segments. It is evident that the items in the second list are features of the visual scene, corresponding to objects and characteristics that exist in the external world and are presented internally as the representational content of the relevant mental states. In other words, this second list represents what a person sees and the kind of information her visual system is processing.

Let us first deal with the list that contains physical (biological) entities, which is analog to the list of entities mentioned in the musculoskeletal case. Indeed, applying Craver's questions to explore if mechanistic levels are involved yields a result similar to the musculoskeletal case: Here, too, several levels of mechanisms can be described in explaining how an individual perceives a specific object: the eye receives light, converts it into an electrical signal, and transmits it to the brain through the optic nerve. Delving into how light is encoded brings attention to the lens and retinal circuitry, elucidated by the roles of rods, cones, and ganglion cells. Going further, the mechanism of phototransduction, involving the activation of photopigments, explains the entire process at the molecular level. A similar hierarchical description can be crafted for other stages of the mechanism (processing in V1, V2, V3, V4, V5, MT, IT, and so forth). In this case, as in spatial perception and muscle contraction cases, a compositional hierarchy of physical entities corresponds to the working components of the neurobiological mechanisms of visual perception. Consequently, concerning the compatibility of the mechanistic levels of organization with all three cases -the neurobiological mechanism of visual perception, the neurobiological mechanism of mouse navigation, and the muscular mechanism of forearm bending- there doesn't appear to be any serious obstacle in implementing levels of mechanisms in the explanations. However, this holds true only if we exclude anything related to mental content from the explanation. If we choose to include content, then the endeavor becomes significantly more complex.

What is then the problem? Reflecting on the visual perception case, the difficulty lies in finding out what to do with the second list, whose members are informational contents and internal representations. More specifically, the members of the second list are mental symbols that stand for

properties of things in the outer world, such as the color, shape, and position of the visual object⁵⁶ and the representation of the perceptual object itself in a particular visual format. What is certainly evident is that, as long as we are discussing explanations regarding only neurobiological mechanisms, things are straightforward. However, as mental content enters the equation, we find ourselves at a dead-end. How exactly can we correlate it with the mechanistic levels to demonstrate that mechanistic explanation can accommodate cognitive phenomena? There doesn't seem to be such a possibility within the classic framework of mechanistic levels examined earlier.

The problem, as thoroughly discussed in the preceding arguments, can be summarized as follows: constitutive mechanistic explanation relies on the concept of mechanistic levels. The unique nature of cognitive phenomena lies in the presence of mental content. Mental content attributed to the states of a system can be studied through the levels of analysis or realization, while its gradual enrichment and integration are represented through the processing levels. Mechanistic levels are compositional levels and do not capture other notions of levels. Based on all the above, it becomes apparent that the model of constitutive mechanistic explanation, as it is, cannot support cognitive phenomena because the relations that exist between the mechanistic levels do not align with the relations found in cognitive explanations.

However, we cannot overlook that as a framework of neurobiological explanations, it works effectively, and in terms of the adequacy of an explanation and its normative use, it is certainly more apt than alternative frameworks. Are there any possible solutions, perhaps different

⁵⁶ Interestingly, within this list, it can be observed that entities are distributed hierarchically. Some entities are simpler and more elemental, while others exhibit greater complexity and integration, suggesting they originate from or are composed of the former. In an attempt to classify them, the resulting hierarchy unfolds as follows: at the lowest level, we encounter line segments and spots of light and darkness. Progressing to the next higher level, we encounter contrasts, contours, lines, and edges. Ascending further, we find movement, shapes, boundaries, and orientation. At the highest level, we encounter images and objects. There is a constitutive relation between these entities: line segments constitute the elementary building blocks of lines, while lines arise from the contrast of light and dark areas. Lines, edges, boundaries, and orientation collectively give rise to shapes and contours. As we ascend to even higher levels, the representation becomes more complex, incorporating color, movement, and other features, eventually encompassing representational features originating from different sensory modalities and cognitive functions. The noteworthy aspect of this hierarchy is the escalating complexity of the representational content as we progress through these levels. Informational processing becomes more complex and integrated as we ascend to higher processing levels. It seems, then, that another difficulty that should be addressed in integrating all aspects of a cognitive capacity in a unified mechanistic explanation is concerned with capturing this ever-increasing complexity that occurs as we move up levels of processing. Maybe, the classic conception of levels of mechanisms is not enough to capture this aspect regarding phenomena taking place in cognitive systems.

interpretations or modifications of the new-mechanistic account, that can maintain it as a viable suggestion for dealing with cognitive phenomena?

The issue of integrating cognitive and mechanistic explanations within a unified framework has been discussed extensively in recent years without a definitive answer, positive or negative, being reached yet. In a conceptually young field, such an outcome is expected. Cognitive science, and especially its latest development as cognitive neuroscience (Boone & Piccinini, 2016), is a relatively new discipline, and the mechanistic explanatory framework is also a relatively recent conception. In such situations, it is neither easy nor appropriate to provide quick answers, as the exploration of conceptual intersections is still ongoing. In this spirit, the next chapter will critically review the discussion regarding the feasibility of unified cognitive mechanistic explanatory models and will explore the idea of recursive decomposition, which is a substantive element in constitutive explanations and in examining phenomena through different levels. A possible approach to resolving the issue of incompatibility between cognitive explanations and the levels of mechanisms will then be explored.

5.4. SUMMARY AND CONCLUSIONS

This chapter initiated the more specialized investigation of the new-mechanistic framework's applicability to cognitive phenomena, particularly emphasizing the concept of levels. Initially, I examined the unique nature of mental phenomena and compared the explanations of two behaviors regarding biological systems - one motor and one cognitive. Through this comparison, I highlighted intentionality and information processing as the features that distinguish cognitive phenomena from other biological phenomena. I also introduced the vehicle-content distinction and emphasized that a complete explanatory model needs to account for both the physical substrate and the informational and representational content of mental states.

After defining and describing the distinctive characteristics of the mental realm, I turned to how explanations of mental phenomena are provided within the multidisciplinary field of cognitive science. My exploration delved into the multiple perspectives and notions of levels used in cognitive science research, including levels of organization, processing, analysis, description, and realization. I used visual perception as a case study to demonstrate how different level notions can be combined to explain a phenomenon. In this case, I utilized Marr's trilevel account, adapting it to

three levels of analysis, three stages of processing, and three levels of organization. This approach aimed to show how the level notions I presented can potentially collaborate in a multi-level and multi-faceted explanation.

Having completed this step, the groundwork was laid to address the main question of this chapter: Is the mechanistic concept of levels, as proposed mainly by Craver (2007), compatible with the multiple notions of levels encountered within cognitive science research? My investigation revealed that the model of constitutive mechanistic explanation, in its current form, cannot adequately support cognitive phenomena because the relations that exist between mechanistic levels do not align with those found in cognitive explanations. The challenge of incorporating mental content into levels of mechanisms is identified as a significant obstacle in applying the new-mechanistic framework to cognitive explanations. Any successful explanatory framework for cognitive phenomena must be able to account for both aspects simultaneously. Moreover, while the mechanistic approach works well for neurobiological explanations, it falls short in capturing the full complexity of cognitive processes. Further exploration and potential modifications to the new-mechanistic account may be necessary to develop a unified framework that can effectively explain both the physical and informational aspects of mental phenomena.

Any conclusions drawn regarding the applicability of constitutive mechanistic explanation to cognitive phenomena concern either its inability to account for the unique nature of mental phenomena or its treatment of levels and hierarchy. Regarding the former, a serious incompatibility between the mechanistic levels framework and mental content has been demonstrated. While mechanistic levels effectively capture the physical and compositional aspects of neurobiological systems, they struggle to incorporate the representational and informational aspects of mental states. There may be potential for expanding or modifying the mechanistic framework to better accommodate the unique aspects of mental processes. This could be made possible by developing a new conceptual framework to bridge the gap between mechanistic explanations and cognitive phenomena. Achieving multi-disciplinary integration between the sciences and methodologies that constitute cognitive science could also play a significant role in developing explanatory models that can capture the full complexity of mental phenomena across multiple levels of description and analysis.

Finally, regarding the latter, this chapter demonstrated the limitations of compositional hierarchy on which mechanistic levels are based, as they prove inadequate in capturing the full spectrum of

level notions used in cognitive science. This limitation highlights the need for a more flexible framework that can accommodate various types of level relations beyond mere part-whole compositions. The integration of multiple level concepts is challenging, and the various notions of levels employed within cognitive science research are not easily reconcilable within the current mechanistic framework, posing significant difficulties in developing a unified explanatory model. A tension between local and global levels is also detected. The local nature of mechanistic levels, while beneficial for specific neurobiological explanations, may be too restrictive for cognitive phenomena that often require consideration of more global, system-wide properties and processes. In any case, there is a need for further research and conceptual development to create a more comprehensive explanatory framework that can bridge the gap between mechanistic and cognitive explanations.

CHAPTER SIX - COGNITIVE MECHANISMS: SO CLOSE, YET SO FAR

6.0. OUTLINE

This is the final chapter of the dissertation, and through it, I explore whether combining psychological with mechanistic explanatory models in a unified, integrated explanatory framework is after all viable. In the previous (fifth) chapter, I thoroughly examined the incompatibilities between mechanistic levels and the various notions of levels, traditionally used to describe and explain cognitive phenomena, utilizing data derived from methodologies that collaborate to jointly answer questions posed within the field of cognitive science. Here, I focus on the concept of decomposition which is a central analytical tool used in mechanistic explanations and psychological as well, in the second case as functional decomposition in constructing psychological models of explanation. I approach the main theme of this chapter by investigating how studying information processing systems through recursive decomposition works. The aim is to arrive, if possible, at a solution that combines psychological and mechanistic models without violating the requirements imposed by the unique nature of cognitive systems and phenomena. I begin discussing the viability of applying mechanistic explanation to cognitive phenomena (6.1. *The Debatable Viability of Mechanistic Explanation in the Cognitive Realm*) by examining various perspectives on integrating psychological and neurobiological explanations, including arguments against the autonomy of psychological explanations (6.1.1.) and proposals for complementary (6.1.2.) or unified (6.1.3.) neurocognitive mechanistic models, the latter expanding on Marr's trilevel account as a potential framework for cognitive-mechanistic explanations. The next section addresses the challenge of incorporating mental properties into mechanistic hierarchies, analyzing different interpretations of the relation between functional and structural aspects in explanatory models (6.2. *Is There Any Place for Mental Properties in Mechanistic Hierarchies?*). In the third part of the chapter, I introduce the concept of recursive decomposition (6.3. *Recursive Decomposition as a Key to Multilevel Cognitive Explanations*) as applied to information processing systems (6.3.1.), and present a case study in visual perception to demonstrate how this approach can be used to analyze cognitive phenomena (6.3.2.). In the final part, I set out my proposal for a framework that could combine psychological and mechanistic explanations in a nearly integrated form (6.4. *Towards "Cognitive*

Rich” *Mechanistic Hierarchies in Explanatory Models*). In this hypothesis I combine elements from recursive decomposition in information processing systems, Marr's trilevel account, and Craver's constitutive mechanistic explanation to create a more comprehensive model for explaining cognitive phenomena.

Note: It is important to note here, that this proposal is made with the reservation of the decomposability of the information processing systems to be explained. As has been discussed multiple times at various points in the dissertation, the "freedom" that near-decomposability provides in the use of decomposition in many complex systems allows for the formulation of such solutions even in controversial cases, as long as a balance between idealization and reality is maintained. Nevertheless, a thorough examination of this issue is certainly necessary, as well as inevitably extensive, and thus could not be included in the present inquiry to the extent that the author deems necessary for a more definitive answer regarding the applicability of mechanistic explanation to mental phenomena.

6.1. THE DEBATABLE VIABILITY OF MECHANISTIC EXPLANATION IN THE COGNITIVE REALM

There is no doubt that the question of whether the new-mechanistic framework for explanation is applicable in cognitive science has engaged the attention of those involved with mechanisms. This inquiry was explicitly posed by MDC in their (2000) seminal paper: “We suspect that this analysis is applicable to many other sciences, and maybe even to cognitive or social mechanisms, but we leave this as an open question” (Machamer, Darden, & Craver, 2000, p. 2) and it remains unresolved. Bechtel (2008) has extensively addressed the issue, examining from various angles how scientists investigate mental phenomena and develop mechanistic accounts for them. Drawing from the unique character of the mind that makes explanation through mental mechanisms distinct from constructing explanatory mechanisms for other biological phenomena, he illustrates the development of mental mechanisms using cases such as the explanation of memory and visual processing. There, he examines how phenomenal and mechanistic decomposition is applied. With Bechtel's contribution, it becomes more feasible to deeply understand what is required to apply mechanistic explanation to cognitive phenomena.

The question arises as to whether the mechanistic model of explanation can support fully integrated explanations, that is, explanations where a research question is answered by the interplay of multiple complementary and cooperating assumptions, data, and research methodologies (O'Malley & Soyer, 2012), essentially, explanations that cover the unique scientific profile of cognitive science⁵⁷. Integrating explanations in a "mosaic unity" is indeed one of the primary goals of the mechanistic approach, as advocated by Craver (2007). This integration can be achieved through contributions from various research fields to explain a particular phenomenon (intralevel integration) and by considering multiple levels of mechanisms to construct explanatory hierarchies (interlevel integration). In the realm of cognitive (neuro)science⁵⁸, two main requirements emerge. Firstly, the chosen explanatory model must be capable of accommodating all the necessary aspects and levels of a cognitive explanation. Secondly, it must provide these explanations by describing the underlying mechanisms. A hierarchy of mechanisms that is solely neurobiological falls short of explaining cognitive phenomena entirely. This is because, as we have observed, the unique nature of cognitive phenomena at each involved level must somehow be included in the analysis. While this requirement may appear to present significant obstacles in the present dissertation, for some, achieving this goal is feasible. Various conflicting views and stances emerge on this question, along with different interpretations of the fundamental factors that could

⁵⁷ According to O'Malley and Soyer (2012), integration is how scientific practice operates today, where various methods and methodologies, datasets, and explanations are recruited to achieve understanding in the subjects studied by the life sciences. They define integration by distinguishing three types, as follows: (1) Methodological integration: "This sort of integration involves directing a range of methods at a particular biological system or research problem in order to gain a multidimensional understanding of how the system works." (p. 60) (2) Data integration: "Often conceived as a problem or at least a major challenge, data integration is the activity of making comparable different data types from a huge variety of potentially inconsistent sources." (p. 61) (3) Explanatory integration: "Refers to both the synthesis of different explanations as well as the import of explanatory (and predictive) models from other research into a specified domain of inquiry." (p. 61)

⁵⁸ The term "cognitive neuroscience," used here instead of "cognitive science," refers to recent developments in the field that reflect a turn to more integrated cognitive neurobiological models for explaining mental phenomena. This position on cognitive science is supported by Boone and Piccinini (2016). They refer to a recent "cognitive neuroscience revolution" and highlight the current state and direction of the field. In support of their view, they note that neuroimaging techniques, such as fMRI, have become central tools in cognitive science research, while computational neuroscience and neural network modeling have become mainstream. Additionally, new interdisciplinary fields bridging cognitive science and neuroscience have emerged and grown rapidly. However, opposing views exist (Núñez, et al., 2019): The authors, in their study, argue that their investigation in author affiliations in the journal *Cognitive Science*, citation patterns, doctoral training of cognitive science faculty, and undergraduate cognitive science curricula, consistently revealed an overrepresentation of psychology. Based on these findings, Núñez et al. argue that cognitive science is currently failing to transition from a multidisciplinary effort into a mature, integrated, interdisciplinary field.

shape integrated cognitive mechanisms. From one perspective, the applicability of the new-mechanistic framework in cognitive phenomena stands as a viable option. This is contingent upon a reconceptualization of how we understand the application of the mechanistic framework within the cognitive realm.

In the following discussion, I will examine approaches that investigate through various perspectives the feasibility of integrated cognitive-mechanistic explanations, that is, explanations where levels related to psychological properties and levels related to neurobiological properties stand together in integrated mechanistic models. Initially, I will explore arguments presented by Piccinini and Craver (2011) and others who disagree with the autonomy of psychological explanations and support the position that, under specific conditions and conceptual considerations, the new-mechanistic framework can effectively accommodate unified cognitive-neurobiological explanations delving into phenomena within the realm of cognitive science. The idea that the analysis of cognitive capacities, which corresponds to functional analysis, is not autonomous but constitutes a mechanism schema that will be transformed into a complete mechanistic explanation once specific missing (structural) details are filled in is a quite popular solution among a number of new-mechanistic philosophers like Craver (2007), Piccinini and Craver (2011) and Boone & Piccinini (2016), while Milkowski (2016a) additionally draws attention to the role that psychological evidence plays in constraining the space of explanatory models.

Albeit promising, this solution also presents several problems. The idea of treating cognitive models as incomplete mechanistic models that will be completed or corrected when supplemented with neurobiological details and, more generally, the idea of seamless complementarity between neurobiological and cognitive explanations is criticized by Stinson (2016) and Stinson & Sullivan (2017) who advocate an intermediate and more compromised position that endorses the possibility of unified neurocognitive mechanistic models, albeit in a non-seamless form. They propose that while a seamless integration of cognitive models and neural mechanisms seems not attainable, a more practical approach involves partial, patchy, and incomplete integration. Lotem and Shagrir (2018; 2021) explore the problems arising from the duality of levels in cognitive phenomena instantiated in neurobiological systems, but they also suggest viewing functional (or computational) and structural properties as complementary rather than autonomous for the development of integrated mechanistic explanations.

The Marr-based critique of Rusanen and Lappi (2007; 2015) against the attempt to unify cognitive and mechanistic explanations will be discussed lastly, along with Zednik's (2017) interpretation of

Marr's trilevel account as a precursor of cognitive-mechanistic explanations, an idea that has the potential to be very promising. Is there any way that Marr's account would supplement levels of mechanisms to form a mechanistic model that could sufficiently accommodate the different notions of levels encountered in cognitive neuroscience research?

6.1.1 Against the Autonomy of Psychological Explanations

Piccinini and Craver (2011) argue that functional analyses of the cognitive capacities exhibited by an information processing system are, in essence, incomplete mechanistic explanations. They propose that integrating functional analyses with neurobiological mechanistic explanations into cognitive-mechanistic explanations is a viable option if we consider the first as preliminary sketches of mechanisms, which will evolve into complete mechanistic explanations once missing structural aspects are incorporated⁵⁹. Nevertheless, functional analyses and their implementation, i.e. componential analysis, were, for Cummins, conceptually distinct, which leads to the thesis that psychological explanations are autonomous. So, how can we start from an analysis that does not refer to any implementing system and arrive at a mechanistic model consisting of functional and structural aspects?

The authors introduce contemporary views of mechanistic explanation and argue that each type of functional analysis is a mechanism sketch, which they further argue is just a kind of mechanistic explanation. They define functional analysis as "the analysis of a capacity in terms of the functional properties of a system and their organization" (Piccinini & Craver, 2011, p. 286) and explain that within this interpretation, at least three types of functional analysis in psychology can be distinguished: task analysis, functional analysis by internal states and boxology. Each can be considered as an elliptical description of a mechanism, that is, a mechanism sketch. To defend this particular claim, they will have to reject the received view of the relationship between functional analysis and mechanistic explanation, which states that these two are distinct and autonomous from one another. They claim instead, that a complete constitutive explanation of a phenomenon based on functional properties should adhere to the limitations set by the structures embodying

⁵⁹ As previously mentioned, functional analysis in psychology and biology was first introduced by Robert Cummins (1975) who stated that "a large part of the psychologist's job is to explain how the complex behavioral capacities of organisms are acquired and how they are exercised" and "the problem of performance resolves itself into the problem of how the analyzing capacities are exercised" (Cummins, 1975, p. 761).

those functional properties. According to them, functional analysis and mechanistic explanation are mutually constrained. They draw from the fact that mechanistic explanation is based on the identification of spatially and temporally organized entities and activities, where each entity or component participates according to its functional properties, and conversely, the existence of specific components plays a determinative role in the possible functions that the system can exhibit. Thus, functions and structures within a mechanism constrain each other. It follows, therefore, that the functional analysis of a system ultimately is not independent of its structural analysis. Furthermore, both explanatory strategies, i.e., mechanistic explanation and functional analysis, aim to reveal the causal structure of a system, but implementation is necessary for an explanatory model because the system's components are the bearers of causal powers and, therefore, they are necessary for understanding causal relations.

Let's see, then, how each of the three cases they distinguish is addressed. Starting with task analysis, they state that it is a way to explore how a number of a system's sub-capacities interact, are organized, and combine to give rise to the overall capacity of a whole system. For the authors, this process is a step in the mechanistic explanation of this capacity because it partitions the phenomenon to be explained into intelligible units of activity, thus placing direct constraints on the function and organization of the structural units that will accomplish the task. This holds true even in cases where the modeling concerns only the flow of information through a system and not how it processes information because even in these cases, any structures that can possibly process the information will be directly constrained. Thus, task analyses can serve as mechanism sketches. The second case they refer to involves explaining cognitive capacities through the study of the system's internal states that mediate between inputs and outputs, which, in the case of the mind, concern propositional attitudes and perceptual states. Explanations of this type fall within the philosophical framework of functionalism. There are two ways to interpret internal states. The authors emphasize that if these states are considered system-level states, they are not internal. On the contrary, the states of a system's components can be considered internal because they are inside the system and, furthermore, they contribute to explaining the global system's state. Considering internal states as component states leaves two options for how one perceives their role in functional analysis: either they are structural components, which aligns with the claim that functional analysis is a sketch of mechanistic explanation, or they are functionally individuated components. The latter is equivalent to black boxes, bringing to the discussion the third case that concerns interpreting functional analysis as boxology. But even if we consider internal states as

functionally individuated states, the interdependency between functional and structural properties of a mechanistic explanation still holds. Consequently, this case also amounts to an elliptical mechanistic explanation. In other words, the principle of multiple realizability allows for different arrangements of structural components, but there must certainly be structural components, and they will be constrained by the functional properties of the system. Thus, in each case, we have everything needed to construct a mechanism sketch that will eventually evolve into a full-blown mechanistic explanation. They conclude: "...there is no functional analysis that is distinct and autonomous from mechanistic explanation because to describe an item functionally is, ipso facto, to describe its contribution to a mechanism" (Piccinini & Craver, 2011, p. 307).

This idea by Piccinini and Craver (2011) is a more detailed and in-depth exploration of Craver's (2007) original proposal, where he developed his mechanistic account based, as he admits, on Cummins' functional analysis. For instance, he mentions "I show how Cummins's account of functional analysis (...) can be supplemented and transformed to become an account of mechanistic explanation" (p. 111). Elsewhere, he notes that "...functional analysis provides an appropriate starting place for constructing an adequate account of mechanistic explanation," and next, "I show how the basic structure of Cummins's account would have to be elaborated and transformed to provide a normatively adequate account of mechanistic explanation" (p. 122). Craver transforms functional analysis into a mechanistic explanation by adding details of the mechanism that realizes the functions. Then, a "how-possibly model" or a "mechanism sketch" can develop into a complete mechanistic model. We see, therefore, here that the idea developed by Piccinini and Craver has already been explored to some extent, although without a specific focus on cognitive mechanisms.

In the same spirit against the autonomy of cognitive/psychological explanations is Milkowski's (2016a) contribution to the discussion, who also advocates that theoretical integration in cognitive (neuro)science can be achieved through the mechanistic perspective but proposes a "slight modification" in Craver's account of "mosaic unity of neuroscience"⁶⁰. One of Craver's (2007) fundamental claims concerns the theoretical integration in neuroscience, and with his account of

⁶⁰ Two types of integration can lead to the Mosaic unity of neuroscience: Intralevel integration, where contributions from multiple research fields work as constraints on the space of possible mechanisms, thereby delineating the plausible mechanisms that explain a phenomenon. And interlevel Integration, where hierarchies of mechanisms that span multiple levels impose constraints on various possible multilevel mechanisms. Alternatively, different research fields targeting different levels, collaborate to link levels in a unified mechanism (Craver, 2007).

constitutive mechanistic explanations through levels of mechanisms, he paved the way for an in-depth and wide-ranging discussion. Craver considers the mosaic unity of neuroscience as a better method for unification compared to reduction and suggests that this can be achieved if research results from different fields are utilized as constraints in a multilevel mechanistic explanation. These fields are considered different and autonomous because “they have different central problems, use different techniques, have different theoretical vocabularies, and make different background assumptions,” while they can be unified “because each provides constraints on a mechanistic explanation” (Craver C. F., 2007, p. 231). Milkowski proposes that the constraints do not operate upon the space of possible physical mechanisms, but they operate on scientific representations (descriptions of mechanisms), with the first corresponding to ontic and the second to epistemic accounts of explanation. Through this differentiation, he argues that the role played by specific scientific evidence in shaping integrated explanations is different. In Craver's account, two models are brought together to form a more comprehensive mechanism. In his interpretation, a scientific field such as psychology serves as a constraint that alters a neuroscientific representation of a particular phenomenon to some extent. In other words, the psychological aspect of the phenomenon isn't necessarily represented in the model, but rather, it influences the model by constraining its representation.

Integrating cognitive with neurobiological explanations is also strongly advocated by Boone and Piccinini (2016) in a very influential work, where they emphasize that multilevel neurocognitive mechanisms are not merely a possibility but a reality emerging from the evolution of cognitive science into cognitive neuroscience. They point out the replacement of traditional cognitive science, which separated psychological from neuroscientific research, with cognitive neuroscience, where the traditional and widely accepted autonomy assumption was proven wrong and abandoned. In other words, the very reality of scientific practice has sealed the union of cognitive and neurobiological explanations into unified models, and this should eventually be acknowledged by philosophical theory.

The authors argue that cognitive science, in its effort to oppose behaviorism, assumed the existence of internal states and adopted the analogy of the mind as a computer. However, this cognition-computation analogy came to imply that research on cognition is independent of research on the nervous system. In fact, though, there was no real basis for this assumption; it was made out of sheer momentum. Behind the “autonomistic picture of cognitive science” lies the idea that “functions can be understood independently from the structures that perform them” (Boone &

Piccinini, 2016, p. 1520). The result was a “two-level picture” in cognitive science, where a *functional level* includes the cognitive, representational, and computational levels, and a *structural level* is non-cognitive, mechanistic, and implementational. However, today, with the prevalence of cognitive neuroscience, research questions about cognitive phenomena are being answered by methodologies that target many levels of organization, and multilevel (constitutive) explanations dominate the field.

So, how does the psychological element enter into these explanations? Since cognitive neuroscience deals with mechanisms that perform computations over representations, the authors propose an account of computation and representation that consists of three propositions through which they connect the functional with the structural element in multilevel neurocognitive mechanisms. Firstly, referring primarily to Dretske’s (1981) informational theory of content, they note that “a vehicle carries semantic information about a source just in case it reliably correlates with the states of the source” (Boone & Piccinini, 2016, p. 1516). Information carrying is further associated with representing: “A vehicle represents a source just in case it has the function of carrying information about the source,” and “for a vehicle to have such a function, the information it carries must be used by some part of the system in which it is embedded. The information is used by the system to the extent that it’s causally relevant to other operations of the system” (Boone & Piccinini, 2016, p. 1516). They conclude with a third proposition regarding computation where they note that “a system performs computations just in case it manipulates vehicles in accordance with rules that are sensitive to inputs and internal states and are defined in terms of differences between different portions of the vehicles it manipulates” (p. 1516). Essentially, what they stress is that since there is this close relationship between the vehicle⁶¹ and content, that is, since the vehicle can both represent and perform computations closely related to the content, the description of neural mechanisms at various levels includes the processing of the content. With vehicles, they refer to the related neural states. Therefore, the neural mechanisms that concern cognitive processes, namely information processing, are constitutive cognitive neural mechanisms all the way down. They bolster their argument by emphasizing that each level in a constitutive mechanism is both structural and functional because it contains structures performing functions. Therefore, in each case, structures constrain functions and vice versa. Thus, since the functional element includes the representational, computational, and cognitive levels, each neural level in the hierarchy models

61

these levels, too, until it reaches a lower point where only neurobiological elements can be detected, and cognitive elements are no longer discernible. Integration is achieved. Is it?

6.1.2. Complementarity, but not Seamless Integration

The answer isn't affirmative for everyone. The interplay of psychology with the neurosciences is at the core of the cognitive science endeavor, and for this reason, successful integration in cognitive neuroscience is a priori desirable. However, there are aspects that make the mapping of psychological entities to neurobiological ones evidently difficult. These difficulties are highlighted by Stinson (2016), Stinson and Sullivan (2017) and Lotem and Shagrir (2018; 2021). Next, we will follow their arguments.

Stinson's (2016) critique, drawing from the integration accounts of Bechtel (2008) and Piccinini and Craver (2011), challenges the notion that cognitive models are an incomplete and temporary form of explanation requiring the addition of neurobiological details to be completed and fully serve its purpose. In the case of Piccinini and Craver, this pertains to mechanism sketches that become full-fledged models when the missing details are filled in, while for Bechtel, the completion process of the cognitive model, derived from the decomposition of a cognitive capacity, is achieved through the localization of cognitive components into neural ones. In both accounts, the success of the integrated model hinges on the incorporation of neurobiological entities. The entire idea harbors a common assumption. Articulated by Piccinini and Craver as "psychological explanations describe aspects of the same multilevel neural mechanism that neuroscientists study" (Stinson, 2016, p. 6), it implies that cognitive and neural mechanisms are seamlessly connected and belong to the same hierarchy. Bechtel "assumes a shared structure between cognitive and neural parts" (ibid). Both refer to the seamlessness that should characterize integrated cognitive-mechanistic explanations, suggesting that a cognitive model is a facet, an incomplete form, or a precursor of a cognitive neurobiological model. Stinson emphasizes that while decomposition and localization are useful heuristics for developing mechanistic models, this does not necessarily mean that they have a normative role in mechanistic integration. The interpretation of what constitutes an integrated model should be more flexible because, in reality, the combination of cognitive and neurobiological entities in a model is not seamless. As she notably states, "I describe a picture where explanations in the cognitive and neural sciences fit together not like the tiles of Craver's mosaic, but instead, like a garment held together by many overlapping patches (p. 3)."

In her critique, Stinson focuses on two main points. Firstly, she demonstrates that cognitive models cannot be regarded or treated merely as sketches of mechanisms. Secondly, she advocates for an alternative version of integration that is not seamless and recognizes the value of cognitive models in formulating explanations as something beyond a temporary and incomplete mechanistic schema. Her primary strategy, strongly based on scientific research in the relevant fields, unfolds through the comparison of two case studies, revealing what truly emerges as a conclusion from the way a cognitive model evolves as opposed to a pure neurobiological case. Thus, she presents a neurobiological model and its development, contrasting it with the evolution of a cognitive model within the domain of cognitive psychology.

The neurobiological case pertains to the development of sodium channel gating models. Through the addition of more and more details, the initial functional and structural descriptions were confirmed, modeled, and gradually began to become more precise in the process of developing initially how-possibly, then how-plausibly, and lastly, how-actually mechanistic models as argued by Piccinini and Craver. So far, so good. However, the evolution of the cognitive case took a very different path. The cognitive model that Stinson used to support her conclusions is Broadbent's attention filter⁶², a cognitive psychology research project. This model had the form of a flowchart and when initially described by Broadbent in 1958⁶³, it did not include details of its physical implementation, although the compatibility of his model with physiological details was undoubtedly desirable. The evolution of the model, however, did not follow the path described by Bechtel and Piccinini & Craver; as Stinson notes, it was considered "an explanatory model in its own right even without the physiological details" (p. 12). This holds true for the subsequent models as well, where any refinements and enhancements were not directed towards detailing the physical implementation. Moreover, the integration of the attention filter models did not incorporate new neurobiological data but rather involved their association with various psychological models dedicated to explaining other cognitive capacities. In no case, however, did this model serve as "a placeholder for a to-be-determined neural model," as Stinson concludes.

⁶² The attention filter model explains results from dichotic listening experiments and the quite infamous "cocktail party effect" where the brain directs a person's focus toward a specific stimulus, typically one that is heard and eliminates from the acoustic scene any other stimulus that could interfere with hearing.

⁶³ Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.

At no means does this imply that the attention filter model remained entirely disconnected from neurobiological developments. As Stinson argues, it served as a starting point for investigating neural mechanisms, a common strategy in cognitive neuroscience. This introduces her second point, which concerns how cognitive and neuroscientific research can ultimately collaborate. The situation she describes is a far cry from a precise correspondence between cognitive descriptions and neurobiological details. Although good cognitive models contain information about the structure of underlying neural mechanisms and certainly guide neuroscientific research, the failure to achieve a complete correspondence does not render the psychological model useless. This is especially true considering its long-term success, validation through psychological experiments, and its harmonious interplay with other psychological models. Therefore, Stinson suggests that if cognitive models include entities that do not map neatly onto physical working parts, the model should not be dismissed. Instead, its ability to accurately describe and predict robust regularities in the world should be considered. In other words, we should allow for the results of neural and cognitive decompositions not to coincide perfectly.

Another point of interest in Stinson's critique pertains to the issue of levels. As has been extensively discussed, the coexistence of two types of entities within a unified model raises concerns, particularly when it is revealed that they form hierarchies that do not fully correspond with one another. Faced with this incompatibility, there are two possible solutions: either to consider the assumption that the two hierarchies should be closely connected as flawed or to hypothesize that there is an error in one of the hierarchies present in our model. According to Stinson, when such a problem arises, the error is attributed to the cognitive model because the neural working parts are too concrete and observable to be disputed. Indeed, as physical entities, neural entities have an ontological priority since, as bearers of causal powers, they serve causality within a system. Thus, they are explanatorily necessary. However, the principle of multiple realizability allows for different configurations of implementations, so the existence of very specific physical entities at the expense of the cognitive model should not be taken for granted since a robust cognitive model could be implemented in other physical entities that also serve its purpose. In any case, Stinson argues that well-established cognitive models that have survived for many years, having been examined and tested through cognitive psychology research, should not be revised in the name of seamless unity of cognitive neuroscience. Besides, as noted by Stinson and Sullivan (2017), the study of the brain spans various levels of complexity, reflecting patterns seen in neuroscience and life sciences. Concepts such as neurons, organisms, and biological systems present challenges due

to their fuzzy boundaries and unexpected behaviors. Despite this, efforts to integrate diverse scientific approaches are promising. By linking results from different experimental paradigms and models, a patchwork understanding emerges, suggesting the potential for an integrated science of the brain.

Lotem and Shagrir have also tackled the issue of the relation between cognitive and neural hierarchies, as we will explore in the ensuing discussion. Similar to the approach taken by Boone and Piccinini (2016), Lotem and Shagrir (2018; 2021) direct their focus toward distinguishing between two levels of description, the functional and the structural, and advocate for the dismissal of the notion of autonomy of the functional level. This is particularly pertinent considering the evolution of research in cognitive science, which has now evolved into cognitive neuroscience. This shift is indicative of a kind of integration between neuroscientific and cognitive research data and methodologies in the explanations provided. Therefore, the question arises: Which framework can encompass this fact and demonstrate how computational-functional properties and structural properties are integrated? Considering that an explanatory model must describe the causal structure underlying a phenomenon - a principle that holds true in mechanistic models - they propose that computational models must be mechanistic to retain their explanatory value. This implies that they must necessarily include information about the implementation of the computation.

However, they emphasize that this endeavor encounters problems such as the following: Firstly, an objection is that the nature of computational explanations is not mechanistic for three reasons. The first pertains to the fact that some computational explanations do not invoke causal relations, which are characteristic of mechanistic explanation, but rather other kinds of principles that ground their explanatory power. The second concerns the fact that computational explanations are not necessarily decompositional. And the third relates to the fact that some computational theories can be formulated in an abstract manner without any information about their implementation. A second objection against integrated cognitive mechanistic models is the ambiguity of how computational explanations may be integrated within a mechanistic hierarchy. This objection has been addressed by those who propose that both levels are integrated at every level of mechanisms, as each mechanism has both functional and structural aspects, and to the extent that the structural aspects have not yet been incorporated, the model is a sketch of a mechanism. However, there are further objections to this response, such as that if recursive decomposition is applied to an explanatory mechanism to produce a hierarchy of mechanisms from a certain level downwards, the

computational level will not be further decomposable. This leads to the uncomfortable situation of ultimately having two different hierarchies, one computational and one implementational, where how the two are related to each other remains unclear.

The authors place significant emphasis on the conundrum of how the two distinct hierarchies relate to each other, particularly when recursive decomposition is applied with the aim of understanding a phenomenon. They propose two solutions for the relation between the two hierarchies. In the first solution, it is posited that a single hierarchy exists that encompasses both computational and implementational properties at every mechanistic level, aligning with the view of computational explanations as sketches of mechanisms. The problem with this solution is that if computational explanations are partial descriptions of the implementational model, then once the implementational properties are eventually added, the computational properties become redundant. Ultimately, there is only one hierarchy, and that is the implementational. However, this is not consistent with scientific practice because computational explanations are regarded as more than mere partial implementational descriptions. If we oppose this conclusion and assert that the computational properties still exist and are nonredundant as they hold explanatory significance, then we necessarily revert to the hypothesis of two different hierarchies that are related to each other through an implementation relation⁶⁴. In this case, because there may not be a perfect match between the two hierarchies, we cannot claim that the implementational hierarchy explains the computational all the way down, but rather that it merely implements it, and from there on, each hierarchy is analyzed separately. Of course, this is not compatible with an integrated cognitive mechanistic model, as it does not involve a part/whole relation between the explanans and the explanandum but only that each hierarchy constrains the other. The implementation hierarchy explains only how the explanandum is implemented. Actually, this means we have resolved the problem of how the hierarchies relate to each other, but we have lost the objective because we no longer have mechanistic integration. In other words, the operation was successful, but the Maharaja is dead.

⁶⁴ The authors here stress that they avoid using the *realization relation* term because they do not want to engage in philosophical debates about the metaphysics of the realization relation.

6.1.3. Employing and Expanding Marr's Trilevel Account

It appears that something more is needed for a cognitive mechanistic explanation. Would Marr's (1982) trilevel account be the missing piece? As we will see, Rusanen and Lappi (2007; 2015) utilize Marr's account as a starting point to evaluate the potential use of the mechanistic framework in explaining cognitive competencies. Respectively, Zednik (2017) advocates the opinion that a reconceptualization of Marr's trilevel account for analyzing information processing systems, incorporating elements from the mechanistic framework, could be a potential solution. Generally, Marr's contribution to the analysis of systems that process information has greatly influenced the debate on the applicability of the mechanistic framework in explaining cognitive phenomena (Rusanen & Lappi, 2007; 2015; Bechtel, 2008; Piccinini & Craver, 2011; Miłkowski, 2016a; Boone & Piccinini, 2016; Lotem & Shagrir, 2021). As we have seen, it includes, in a unified account, the role of the content of a cognitive state, as well as the computational properties that allow the state to participate in a complex network of interactions with other states, serving operations that we typically categorize as perception, thought, decision-making, problem-solving, and more; all of which constitute our cognitive world. At the same time, it explains how the brain represents and manipulates information to achieve these computational goals, as well as how the relevant processes, transformations, and manipulations are implemented in a physical substrate such as the brain. In other words, it is the closest approximation to an integrated cognitive mechanistic explanation.

Rusanen & Lappi (2007; 2015) challenge the applicability of the mechanistic model to the cognitive domain, claiming that while it can bridge neuroscientific explanations across different levels from molecular to systems neuroscience, it cannot account for the "genuinely cognitive level," which they equate with Marr's computational level. They argue that computational mechanisms are special within the mechanistic framework because they involve information-processing tasks that are rational and semantic-sensitive. Specifically, they assert that "the mechanistic model of explanation cannot capture that which is special in explanations in cognitive and neurocognitive sciences generally." (2007, p. 1). Marr's computational level, addressing the "what" and "why" of a system's goal and computation suitability, is essential for grasping the information extraction problem it solves, distinct from the "how" questions tackled by the algorithmic and implementational levels. They further contend that the explanatory role of the computational level cannot be captured by the standard forms of mechanistic explanation. Neuro-cognitive science explanations bifurcate into two modes: one rooted in mechanisms and the other in cognitive competencies, yet current mechanistic

models only address the first. Consequently, the prevailing mechanistic approach falls short of being a comprehensive explanatory model for cognitive-level competencies within neuro-cognitive systems. The authors advocate for an alternative explanatory model that can encapsulate these distinct characteristics, which significantly transcend the scope of mechanistic explanations, particularly at the computational level.

Zednik (2017), following a more positive line of argument, argues that many of the challenges faced today by Marr's account, which is very influential for cognitive explanations, could be bypassed if its scope is expanded and considered as an early version of mechanistic explanation. He contends that the questions posed at each level of Marr's framework are similar to those asked to discover and describe cognitive mechanisms. Zednik aligns Marr's computational level with inquiries about what a mechanism is doing and why, the algorithmic level with questions about how it operates, and the implementational level with questions about where the operations are carried out⁶⁵.

He examines each level more closely. The computational level deals with the cognitive system's computational goals, meaning it focuses on what the system is doing, specifically its behavior. As an information processor, the system's behavior entails mapping sensory inputs to behavioral outputs. However, Zednik argues that this alone does not sufficiently specify the phenomenon to be explained. Questions about why certain computations are beneficial to the overall function of the system are also necessary. The contextuality and appropriateness of a mechanism are integral to a cognitive explanation, complementing the input-output mapping that defines the system's behavior. Zednik concludes that the particular interpretation of Marr's upper level is entirely compatible with mechanistic explanation. He further contends that the intermediate level, i.e., the level of algorithm and representation, pertains to cognitive modeling, describing how inputs are transformed into outputs. He aligns this level with functional analysis (Cummins, 1975). He also notes, in favor of the idea also supported by Piccinini and Craver (2011), that functional analyses are incomplete

⁶⁵ Zednik's interpretation of Marr's levels of analysis reminds Bechtel's (2008) distinction between phenomenal and mechanistic decomposition. In phenomenal decomposition, an explanandum phenomenon, such as memory, is analyzed into the phenomena that constitute it, namely, memory phenomena. Similarly, in mechanistic decomposition, which is divided into functional and structural decomposition, the focus is on the operations performed by the systems responsible for the memory phenomena (functional decomposition) and the component parts that carry out these operations (structural decomposition). Bringing all these aspects together entails constructing a mechanistic explanation. Therefore, when considering mental mechanisms, Marr's account already offers many parallels that render it highly relevant to the mechanistic explanatory framework.

mechanistic models, i.e., mechanism sketches, and emphasizes the importance of the “model-to-mechanism mapping constraint (3M)”⁶⁶ for assessing whether a cognitive model accurately describes the mechanisms that answer how-questions. The last level to consider is the implementational level, but, as Zednik suggests, its complementarity to the other two levels is quite challenging. He notes that “it remains unclear what it actually takes to show that the production rules, symbols, equations, and/or variables specified by a cognitive model are realized by complex and potentially unruly physical systems such as the brain” (Zednik, 2017, p. 395). To address this challenge, he proposes a solution: to view implementational features as the missing details that, once uncovered, will transition a cognitive model from a mere mechanism sketch to a fully developed mechanistic explanation. And with this idea, he aligns with all the previously mentioned authors who argue that functional analyses are incomplete models of mechanisms.

Zednik, after all, aligns with the spirit of Piccinini and Craver’s work, as he too regards cognitive models “which involve(s) *functionally analyzing* the complex behavioral or cognitive capacity being explained into an organized collection of simpler capacities” (Piccinini & Craver, 2011, p. 392) as incomplete mechanism sketches and attempts to determine the kinds of details that are left out and should be filled in. Zednik’s divergence from Piccinini and Craver lies in placing his idea within Marr’s trilevel approach, mapping the cognitive competence to be explained onto the computational level, the mechanism sketch onto the algorithmic level, and the details that complete the mechanism onto the implementational level. This maneuver, I appreciate, affords greater scope for accommodating the distinctive nature of the mental, in the sense that it differentiates it from the analysis of psychological/cognitive capacities, which pertain solely to Marr’s algorithmic and representational level.

Zednik is not alone in this effort to utilize Marr’s ideas for cognitive-mechanistic integration; Bechtel (2008) also advocated Marr’s view as a means to combine cognitive and mechanistic explanations. Bechtel posits that Marr’s trilevel approach essentially provides what was missing in mechanisms to account for cognitive phenomena, which involves analyzing a problem as an information processing task. In his discussion on information processing mechanisms, Bechtel

⁶⁶ The 3M constraint is about successful explanatory models, where “(a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism.” (Kaplan & Craver, 2011, p. 611)

refers to the distinction between content and vehicle, stating: “What is critical to understanding an information processing mechanism is how content is changed as a result of causal operations that change vehicles, for it is in virtue of content that the mechanism is linked to the states outside the mechanism” (2008, p. 25). Therefore, vehicle and content are inextricably linked. Does the study of the vehicle suffice to understand mental mechanisms? Marr (1982) considered it insufficient to analyze a system based solely on the causal operations performed on the vehicle, as the analysis of content, independently from its implementation, also includes elements that will lead us to understand the mechanism. He proposed a different perspective for studying the tasks performed by information processing systems, which he named *computational theory*, corresponding to what is widely known as Marr’s computational level of analysis. Bechtel separates Marr’s levels of analysis into two aspects: one, being outside the mechanism, relates to what the mechanism does and why, namely “the mechanism in context,” and corresponds to the computational level. The other two levels (representation and algorithm and hardware implementation) concern the inside of the mechanism, that is, how and where the physical processes occur. Importantly, as Bechtel mentions, Marr considered his three perspectives or levels to be coupled, but only loosely. In other words, Marr suggested a constraint and not a determination relation between the levels. From the way Bechtel interprets and conveys Marr’s ideas, we can conclude that he views Marr’s account as a key to understanding how informational and physical processes are connected and that he considers informational processes not to be exactly part of the mechanism itself but an external constraint, which links the content with the physical processes that constitute the mechanism. Such an interpretation of Marr’s account is more compatible with the loose connection between cognitive models and neurobiological mechanisms proposed by Stinson (2016).

6.2. IS THERE ANY PLACE FOR MENTAL PROPERTIES IN MECHANISTIC HIERARCHIES?

It is now time to revisit the former discussion with a critical eye and explore whether there is potential for a fruitful coexistence of the mental (or informational⁶⁷) and the neurobiological aspects of a phenomenon within a unified mechanistic model and in what form such an interplay could be conceptualized. But let us first start with some observations about the terminology used, the entities involved, and what issues the different approaches focus on. As we observed, some scholars treat psychological explanations as *functional analyses* (Craver C. F., 2007; Piccinini & Craver, 2011). Others position them within a broader *functional level*, which encompasses cognitive, representational, and computational aspects - distinct from the *structural level*, which is non-cognitive, mechanistic, and implementational (Boone & Piccinini, 2016; Lotem & Shagrir, 2018; Lotem & Shagrir, 2021). Additionally, some consider psychological explanations as the fundamental focus of models for cognitive phenomena (Stinson, 2016; Stinson & Sullivan, 2017), while others view them as corresponding to computational theories or Marr's computational level (Bechtel, 2008; Rusanen & Lappi, 2007; Rusanen & Lappi, 2015; Zednik, 2017). In many discussions, the term *information processing* (inherited from the analogy between the mind and the program running on a computer) persists (Bechtel, 2008; Marr, 1982), along with the distinction between *content* and *vehicle* (Bechtel, 2008). Not all terms related to mental properties are synonymous, and this reflects the conceptual difficulties inherent in precisely characterizing mental entities and processes. Additionally, various metaphysical theories about the nature of the mind come into play. Regardless of the terminology used, it is widely accepted that two distinct hierarchies of entities (or properties) are involved in explaining mental phenomena: one pertaining to physical aspects and the other to mental (or informational⁶⁸) aspects.

⁶⁷ In discussions related to the explanation of mental phenomena, we often encounter both the term "mental" and the term "informational." The use of these two terms is often accompanied by various theoretical implications. For example, Palmer and Kimchi (1986), in their extensive review of the Information processing approach to cognition, use the term mental for the non-decomposed level of description associated with empirical reality while they use the term informational for the lower levels derived from the decomposition of the upper level that is applied for explanatory purposes. That is, the second term has a more general reference. With the same rationale, both terms are used here, in the present investigation.

⁶⁸ Or *cognitive* or *psychological* aspects. In discussions regarding explanations of mental phenomena, we often encounter these terms too, in the same role, more or less, as the terms *mental* and *informational* are used.

In the discussion, we encountered two primary stances regarding the feasibility of fully integrating mental properties into a mechanistic model. These stances concern the autonomy of psychological explanations. A common point of contention lies in whether psychological explanations, i.e., explanations referring to mental properties, can be fully autonomous or if they are ultimately grounded in physical implementations. Scholars such as Piccinini and Craver (2011), Milkowski (2016a), and Boone and Piccinini (2016) have debated the issue of autonomy. They argue that the distinction between structural and functional aspects within an explanatory model is somewhat artificial and has become established due to historical precedents. Every structural property constrains functional properties, and vice versa. Therefore, it doesn't make sense to examine them separately. On the other side, scholars like Stinson (2016), Stinson and Sullivan (2017), and Lotem and Shagrir (2018; 2021) while they consent with the idea of a close relation between both kinds of properties, physical and mental, they disagree with the idea to consider them together and to end up with an explanation that is reduced to the mechanistic and implementational aspects of the explanandum. As revealed in the forgoing discussion, achieving complete integration is impossible without sacrificing the autonomy of psychological explanations. Unfortunately, as we have seen, this compromise also undermines their explanatory utility. Moreover, considering that functionalism essentially implies the autonomy of functional explanations, starting from functional analysis and ending up denying autonomy is akin to endorsing functionalism without functionalism. Therefore, Stinson's solution for partial integration offers a convenient compromise that avoids throwing the baby out with the bathwater.

Another issue identified in the accounts that stand against the autonomy of psychological explanations (Craver, 2007; Piccinini & Craver, 2011; Boone & Piccinini, 2016) is that they limit mental properties to their functional roles, disregarding the unique perspective they offer in information processing explanations, as Marr (1982) had already emphasized. Importantly, this motivated him to propose computational theories for understanding mental phenomena. Indeed, treating mental properties based solely on their functional roles makes psychological explanations too narrow to encompass all forms of psychological inquiry in cognitive science. While Cummins' functional analysis provides a valuable and foundational framework for understanding cognitive phenomena, it might not be sufficient on its own to capture the full complexity and richness of cognition. It may oversimplify and not provide enough detail to fully capture the unique nature of different cognitive processes.

This particular problem is detected in the suggestion put forth by Piccinini and Craver (2011), which practically equates psychological explanation with functional analysis. This problem brings us back to Craver's initial formulation of his mechanistic account: If one examines how Craver (2007) formulates his normative proposal regarding what constitutes a mechanistic explanation, it becomes evident that Piccinini and Craver (2011) follow a similar strategy in their proposal for integrating psychology and neuroscience. In particular, in the same line as Craver's, they demonstrate how Cummins' functional analysis can be complemented with implementational details, resulting in the development of a *how-possibly model* or a *mechanism sketch* into a complete mechanistic model. While Craver offered this particular approach for explanations provided by neuroscience, the authors argue that a similar adaptation could apply to integrated cognitive-neuroscientific explanations. In both cases, the analysis of certain functions is assigned to specific structures. But, if a cognitive-neuroscientific explanation and a purely neuroscientific one are treated in the same way, that is, as functional analyses plus implemented details, how do the two explanatory strategies differ? It seems that they do not differ at all.

Moreover, Piccinini and Craver (2011), in supplementing their argumentation for integrating psychology with neuroscience, put functional analysis at work: they employ a case study involving the mechanisms of neurotransmission to demonstrate how structural details complement functional analysis, developing a mechanistic model. They stress, "the same, we claim, extends all the way to the highest levels of the neuroscience hierarchy" (p. 306) and they imply that this example is representative of psychological explanations, too. From this example, one might infer that a similar approach could be applied to explain, let's say, the internal representation of an object's contour - achieved through the combination of regions encoding darkness or brightness. This type of explanation rests on the following reasoning: Functions are decomposed into sub-functions and are ascribed to their implementational details. However, it is widely acknowledged that these two kinds of phenomena, i.e., neurotransmission vs. visual perception, cannot be explained in precisely the same manner: the first is solely neurobiological, and the second is a cognitive phenomenon that takes place in a neurobiological organism. In the latter case, as we have previously demonstrated, an additional layer exists - the processing of content - unfolding across multiple levels of processing and organization. Furthermore, as reviewed in former sections, Stinson (2016) stressed the fundamental difference in developing neurobiological and cognitive models, which further emphasizes the need to treat them as two different explanatory strategies.

Cognitive-neuroscientific explanations are unique because they specifically address cognitive phenomena. Consequently, their explanatory approach should be equally distinctive. A functional analysis alone is insufficient to fully capture the specificity of these explanations when examining a cognitive capacity. To establish a comprehensive hierarchy of cognitive mechanisms, an in-depth analysis of cognitive capacities is necessary. This analysis must be thorough, incorporating both neurobiological and cognitive elements, to provide a complete understanding from the highest to the lowest levels of explanation. Otherwise, we find ourselves with an analysis of cognitive sub-capacities confined to the upper behavioral level and a hierarchy of neurobiological mechanisms that explains solely the implementation of this behavior.

And in the final analysis, the strategy of breaking down complex capacities into simpler building blocks is not unique to psychology. Cummins proposed functional analysis for biological explanations in the first place, and one of his main examples was the function of the heart, which is a biological and not a psychological capacity⁶⁹. Therefore, the claim that psychology relies heavily on functional analysis may not distinguish it from other scientific disciplines, like biology, or may not capture the unique nature of psychological phenomena. The entire treatment of the issue by the authors, as well as the example they present at the end of the article concerning the mechanisms of chemical neurotransmission, might lead one to argue that any biological functional analysis is an elliptical mechanistic explanation. So far, so good. However, there is no mention of how and where intentionality -the essential property of cognition- fits into this framework. And it cannot fit, of course, because, as we have seen in the previous chapter, the relation between mental properties and the hierarchy of explanatory mechanisms is not a mere part-whole one. Consequently, they

⁶⁹ Cummins had indeed stated that functional analysis is an explanatory approach that aligns with psychological explanations. However, as evident from the following quotation, this compatibility is more suited to behavioral psychological explanations rather than psychological explanations that assume and analyze internal states. The latter reflects what cognitive science examines today.

Perhaps the most extensive use of the analytical strategy in science occurs in psychology, for a large part of the psychologist's job is to explain how the complex behavioral capacities of organisms are acquired and how they are exercised. Both goals are greatly facilitated by analysis of the capacities in question, for then acquisition of the analyzed capacity resolves itself into acquisition of the analyzing capacities and the requisite organization, and the problem of performance resolves itself into the problem of how the analyzing capacities are exercised. [...] Watson's analysis proved to be of limited value, but the analytic strategy remains the dominant mode of *explanation in behavioral psychology* [emphasis added]. (Cummins, 1975, p. 761)

cannot be included in the hierarchy of compositional levels within a constitutive explanation without significant modifications of the mechanistic explanatory model.

Compared to Piccinini and Craver's, the approach taken by Piccinini and Boone (2016), at first glance, appears to be broader in terms of the role they assign to cognitive properties. The *two-level picture* of cognitive science that they aim to unify encompasses cognitive, representational, and computational aspects at the functional level. Here, the term "functional" inherently carries more richness than the "functional" used in the previous interpretation of Cummins' functional analysis and of pure functionalist theories of mental states. In essence, within this so-called functional level, the authors place the entire philosophical core of traditional cognitive science, which comprises three philosophical positions: functionalism, computationalism, and representationalism. Piccinini and Boone attempt to argue that functional and structural levels should not be examined in isolation. To achieve this, they propose a triple correlation between vehicle and content, encompassing how a vehicle carries information, represents it, and computes in close correspondence with specific content. Consequently, if the content's intended function is effectively carried out by the vehicle acting on its behalf, the authors conclude that the content is adequately represented in the mechanistic model. Their hypothesis presents challenges, such as the uncertainty regarding whether the correlation they describe is sufficient to fully capture the intricate relationship between cognitive functions and neural mechanisms. Additionally, some cognitive processes may lack a direct neurobiological correlate. Beyond these objections, the crux of the matter lies in the reduction of mental content, the explanatory role of which is identified with the role of the vehicle, that is, its physical implementation, since more or less what is being argued is that anything concerning mental properties is included in the description of physical mechanisms, provided that the former are fully represented by the latter. The authors explicitly oppose the autonomy of the "privileged cognitive level", as they characterize it, and despite their denial of endorsing a reductive view of cognitive science, their treatment of mental content suggests otherwise.

In any case, denying the autonomy of psychological explanations seems to lead to a dead end. One possible solution to advocate for the possibility of genuine cognitive mechanistic explanations would be to explore Marr's account. Marr's approach avoids both endorsing mere functionalism and reducing mental properties to their neurobiological implementation, providing a more nuanced perspective on the interplay between cognitive and neural mechanisms.

Marr (1982), by introducing three levels of analysis, allows mental properties to be incorporated into a model without reducing them solely to functional roles or their physical implementation. This happens because the content is preserved at the upper computational level, the functional role is analyzed at the intermediate algorithmic and representation level, and the physical realization is described at the lower implementation level. These three levels are related through realization, implying that they pertain to the same entity, that is, the system presenting the explanandum mental behavior, yet each level of analysis represents a distinct facet of the phenomenon. The challenge lies in integrating Marr's trilevel account with the levels of mechanisms proposed by Craver (2007) within the framework of constitutive mechanistic explanation, which will be the focus of the next sections.

6.3. RECURSIVE DECOMPOSITION AS A KEY TO MULTILEVEL COGNITIVE EXPLANATIONS

As we have seen, Bechtel (2008) points out that cognitive mechanisms represent a special case because, in addition to processing physical entities, they also regard informational entities. To demonstrate how this combination can be reconciled with the mechanistic account, Bechtel invokes Marr's (1982) trilevel account and the concept of *recursive decomposition* as presented by Palmer and Kimchi (1986) in their thorough review of the Information-Processing Approach to Cognition. The ideas put forth by Marr⁷⁰ and Palmer & Kimchi may have the potential to bridge information processing accounts with explanations of underlying physical processes. Subsequent sections will examine whether this combination is ultimately feasible. We will explore how recursive decomposition, a fundamental principle governing psychological explanations within the

⁷⁰ It should be noted that David Marr's significant contribution to cognitive science, particularly in the study of visual perception, extends beyond his conception of the three levels of analysis in an information processing system. His theory and research program for studying vision (1982) exemplify the computational approach to investigating cognitive phenomena. He was strongly influenced by the development of Artificial Intelligence during his time, which emphasized an engineering approach to exploring mental processes. Simultaneously, advances in neuroscience and their discoveries related to visual mechanisms (such as Hubel and Wiesel's work) guided his research in the field of visual perception up to the present day. Marr's theory exhibits modular characteristics, with many of his proposed vision processes being modular (Gordon, 2004). In this spirit, he identified three successive stages of visual processing. Image formation: This stage represents the spatial distribution of intensity values across the retina. Primal Sketch: The transformation of these intensity values into specific forms of information. 2^{1/2}D Sketch: Relates to orientation and depth of visible surfaces. 3D Model Representation: This concerns a three-dimensional model of the observed object.

framework of Informational Processing Theory, can be used in conjunction with Marr's trilevel account to create an appropriate foundation for structuring cognitive mechanistic explanatory models.

The idea of recursive decomposition is part of the Information Processing Approach to Cognition (Palmer & Kimchi, 1986) and describes how cognitive psychology provides explanations through the lens of computational theory of mind. It exhibits significant commonalities with Marr's three levels (1982), and as I will argue, if combined with the mechanistic account, it could complement a framework for constitutive cognitive-mechanistic explanations. Such an approach might successfully accommodate the two fundamental notions of levels used in cognitive science, levels of organization and analysis/realization while maintaining the explanatory value of psychological explanations for the same reasons that Marr's account does. The advantage of this combined approach is that it accounts for the functional role of mental states (typically included in functional analysis) while also describing how their informational content changes. Since both aspects are necessary for the informational description of a mental event, this perspective is certainly more comprehensive.

It exhibits two additional positive characteristics. First, it doesn't begin with implementation details (the mechanism elements) and then incorporates cognitive aspects later. Instead, it initiates with and relies on cognitive aspects from the outset. While implementation details have ontological priority because they provide causality to the model, cognitive aspects take explanatory precedence. Given the phenomenon's multiple realizability, an informational processing model could be implemented in various ways. For this reason, when explaining a cognitive phenomenon, it is essential to begin the explanation directly from the phenomenon itself rather than implying it indirectly through its implementation. Secondly, the idea of treating psychological explanations as mechanism sketches, which, in a more comprehensive form, will become a complete mechanistic explanation, with all the objections we analyzed that it faces, is not applicable here. According to the proposed combination, an explanation must encompass all three elements: content, functional role, and implementation. None of these can be considered part of the explanatory model's provisional, incomplete form.

Two additional factors make the combination of elements derived from these two accounts a solid foundation for developing a unified model of information processing mechanisms. Let's begin with the first factor, which revolves around the use of recursive decomposition in analyzing information

processing systems. As I've demonstrated here and particularly in the previous chapter during my exploration of explaining cognitive phenomena, a constitutive cognitive-mechanistic explanation should consider cognitive properties at all levels of the hierarchy, not just the top level of observed behavior, which is the level embedded in empirical reality. Otherwise, it is nothing more than an integrated cognitive-neurobiological explanation at the top level, complemented by a hierarchy of purely neurobiological mechanisms below that level. The account proposed by Palmer and Kimchi, incorporating recursive decomposition of informational states and events, establishes a hierarchy of informational operations that explains a cognitive phenomenon at progressively lower levels of composition. The concept of recursiveness is particularly relevant in this endeavor because it illustrates how cognitive explanations can be developed all the way down.

The second factor concerns how a unified model will accommodate the different notions of levels used to analyze a cognitive system. At least two fundamental notions of levels should be represented in such a unified model: (compositional) levels of biological organization, which are responsible for the constitutive character of the mechanistic explanations, and levels of analysis or realization, which reflect the multifaceted ontological nature of mental events. Specifically, the latter refers to the fact that a particular neurobiological state possesses simultaneously physical and informational properties, and this characteristic is unique to cognitive phenomena. These properties represent different aspects of the same entity and stand in a realization relation. This realization relation also applies to the levels where the entities and their processes are described. Following Marr's proposal, the three aspects that should necessarily be represented in a model of cognitive mechanisms to be explanatorily complete are the informational content of the cognitive phenomenon, its functional role as part of a cognitive system, and the structural details that demonstrate how the phenomenon is implemented in a neurobiological system. Combining these three ways of understanding information processing systems with recursive decomposition could provide a useful framework for developing a hierarchy of cognitive mechanisms.

In what follows, we will begin by presenting the use of recursive decomposition in the analysis of information processing systems. Then, we will proceed to build a framework for constitutive cognitive-mechanistic explanations.

6.3.1. Applying Recursive Decomposition to Information Processing Systems

Essential for understanding how a cognitive phenomenon can be explained through a hierarchy of mechanisms, specifically via a constitutive mechanistic explanation, is the need to make clear how a

mental event occurring within an information processing (IP) system can be analyzed within a hierarchy of levels. Each level of the hierarchy describes in greater detail how a cognitive capacity is broken down into sub-capacities. Therefore, we refer to a process of recursively applying functional decomposition to the system, where at each step, the function of the higher level is described and explained.

Functional decomposition involves analyzing a system's activity by breaking it down into sub-functions performed by independent subsystems. These subsystems have their own specific domains of application and operate in an intrinsic way while maintaining minimal interaction between them. It lies at the heart of the traditional computationalist approach in cognitive science and is closely linked to the widely recognized assumption of the modular organization of the mind (Fodor, 1983). According to the latter, certain cognitive processes (particularly those involved in perception and language) are carried out by specialized, encapsulated modules, which are genetically specified, independent, and exhibit distinct functions. These modules operate autonomously, outputting data suitable for central processing while remaining insensitive to central cognitive goals and informationally encapsulated (Robbins, 2017). For many decades, the idea of modularity and the analogy of the mind as a general-purpose problem-solver dominated theories and research methodologies in cognitive science. However, this strictly defined conception of the mind reminiscent of earlier notions such as *phrenology* (proposed by Franz Joseph Gall, 1758-1828), which posited specialized 'centers' in the cerebral hemispheres serving specific intellectual or moral functions, and later *organology* (endorsed by Paul Pierre Broca 1824-1880 among others), which hypothesized various *organs* corresponding to distinct mental functions, faced intense scrutiny from approaches that emphasized the complex character of the mind and mental life, more generally (Richardson, 2001). These alternative approaches highlighted that the mind and brain are not completely decomposable and should be studied as near-decomposable systems (Simon, 1962), that is, systems that possess simultaneously modular and systemic properties. Nevertheless, the idea of functional decomposition continues to offer valuable insights into our understanding of cognitive phenomena. It still influences cognitive modeling and research methodologies in many fields of cognitive science. Therefore, what will be presented from here onward in the sections focusing on the mechanistic modeling of an information processing system presupposes and largely adopts the principle of functional decomposition and the modular organization assumption.

The conception of information processing is also a central element in cognitive science, reflecting fundamental aspects relevant to the practice of cognitive psychology. In essence, cognitive systems are considered information-processing systems and are studied as such within traditional cognitive science. This conception is closely related to and originates from functionalism, drawing an analogy between the human mind and a digital computer. When attempting to explain a mental event, an *Information Processing Theory* is articulated, which, according to Palmer and Kimchi (1986, p. 39), is based on five central assumptions, which are as follows:

- (1) *Informational Description*: Mental events can be functionally described as “informational events”, each of which consists of three parts: the *input information* (what it starts with), the *operation* performed on the input (what gets done to the input), and the *output information* (what it ends up with).
- (2) *Recursive Decomposition*: Any complex (i.e., non-primitive) informational event at one level of description can be specified more fully at a lower level by *decomposing* it into (a) a number of components, each of which is itself an informational event, and (b) the temporal ordering relations among them which specify how the information “flows” through the system of components.
- (3) *Flow Continuity*: All input information required to perform each operation must be available in the output of the operations that flow into it.
- (4) *Flow Dynamics*: No output can be produced by an operation until its input information is available and sufficient additional time has elapsed for it to process that input.
- (5) *Physical Embodiment*: In the dynamic physical system whose behavior is being described as an informational event, information is embodied in states of the system (here called *representations*), and operations that use this information are embodied in changes of state (here called *processes*).

We will focus on assumptions 1, 2, and 5 and start with the assumption of informational description. Every informational event can be described in two complementary ways: as a mapping between input and output and as an operation between them. Informationally, an event is described as a mapping between an input and an output. It is called *mental mapping* if the operation performed between input and output is ascribed to the human mind, and more generally,

concerning any mapping between inputs and outputs, it is called *mapping theory*⁷¹. It is a complex function that matches the incoming to the outgoing information, but without saying anything about what is inside the black box, where the processing takes place, and how this matching is done. It can be decomposed into *mini-mapping theories*, which are functions within a function and map inputs to outputs. Mapping theories regard the upper, undecomposed level of the undecomposed mental event and are directly tied to empirical reality, while mini-mapping theories regard the informational events that are products of the mental event's decomposition. A complete description of a mental event includes addressing the what question, which pertains to the input-output correspondence, and the how question, which involves analyzing the operation performed on the input to yield the specific output. This role pertains to functional description, which concerns how the mind/brain behaves within its environmental context. As Palmer and Kimchi (1986) explain, functional descriptions within the framework of the Information Processing Approach to Cognition are characterizations that include information and operations that relate to information.

The mind, therefore, processes information by receiving it as input from the environment (or other systems) and, through operations, transforms it into a different form of information or behavior in the environment. But how does this transformation of information occur? It happens through processes within the mind, which is generally considered a 'black box.' According to the Information Processing Approach to Cognition, to explain this transformation, an informational event must undergo decomposition into smaller informational events that constitute it. Recursive decomposition can be used to analyze a mental (or, more generally, informational) event in informational events that constitute it, each of which is described as input-operation-output, too. This process will reveal a flowchart representing the organization of the individual operations that interact to produce the explainable process⁷². In this way, a recursive decomposition of a mental event will produce a flow diagram representing at each level the temporally organized operations that constitute the upper-level operation and mini-mappings for each operation. The repeated informational decomposition results in creating a tree graph, and each iteration defines a new, lower level in this hierarchy (see *Figure 14*).

⁷¹ Palmer and Kimchi note that mapping-theories are very similar to Marr's (1982) computational theories.

⁷² Palmer and Kimchi assert that Informational Theories produced by the recursive decomposition of the mental/informational events correspond to Marr's (1982) algorithmic level.

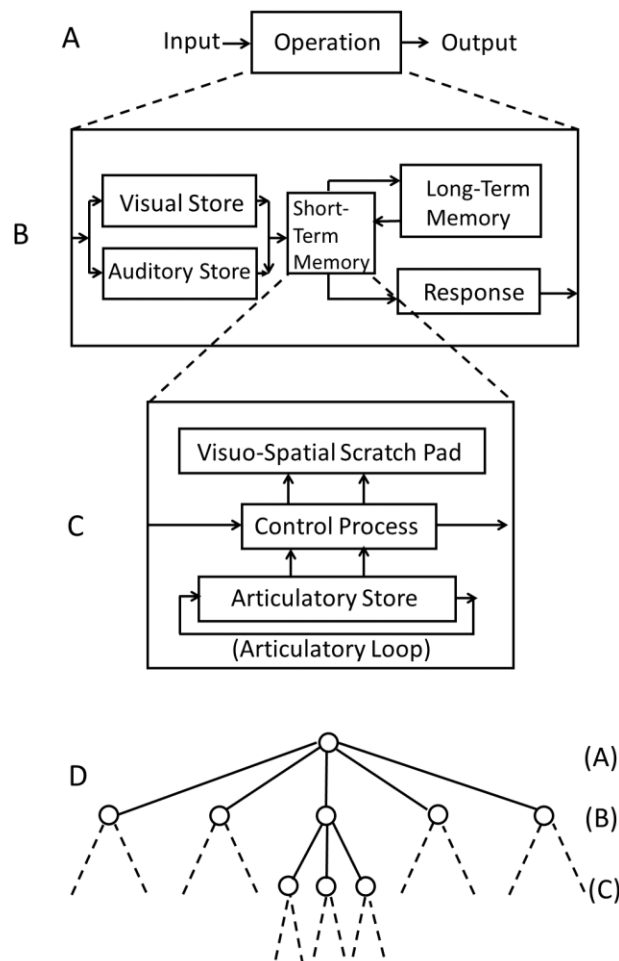


Figure 14: Recursive decomposition of mental processes. The mind can be described as a single, complex informational operation that maps input stimulation to output behavior (A). This operation can be decomposed into an information flow diagram of several simpler component operations (B), each of which can be further decomposed into still simpler operations. The resulting recursive structure of theoretical decomposition can be represented by a hierarchical graph (D) in which the nodes at a given level correspond to the component operations of the flow diagram at that level and the links (arcs) connecting them represent the decomposition relations between different levels of description. Reproduced and improved graphically from Palmer and Kimchi (1986, p. 41).

To illustrate how recursive decomposition would apply in analyzing a certain mental event operating in the mind, Palmer and Kimchi present an example based on Atkinson and Shiffrin's (1968) model for explaining human memory. The renowned *Multi-Store Memory Model* or *Modal Model*, one of the most widely recognized models and still relevant today, serves as a representative example of how the information processing theory addresses human memory. It distinguishes three distinct memory stores, each with different capacities: the *sensory store* (with separate stores for each sense), *short-term memory*, and *long-term memory store*. Various control processes govern the

flow of information exchanged between these stores. In essence, this model posits that when someone attends to information registered in one of the sensory stores (for example, visual or auditory input, as demonstrated by Palmer and Kimchi's exemplary flowchart), part of that information transfers to the short-term memory store. It then undergoes a rehearsal process and subsequently moves into the long-term store. The flowchart used by Palmer and Kimchi depicts how information is stored in memory and the potential generation of relevant behaviors based on this stored information. In their diagram, they depict a further decomposition of one of the component operations, the short-term memory operation, which leads to a third level of description (denoted as C in the diagram) containing three operations that constitute it: the visuospatial *scratch pad*, the *control process*, and the *articulatory store* operations.

At this point, it should be emphasized that the rationale behind explaining a mental event through recursive decomposition into increasingly simpler informational events does not suggest that the black box of the mind opens and what happens inside is revealed. Most of the 'mystery' that governs the original mapping theory of the mental event is transferred to the flow diagram. Essentially, decomposition aims to factor out some part of the complexity that is implicit in a mental operation concerning a mental (informational) event, making it explicit in the flow relations between a number of simpler operations. As the decomposition progresses down to lower and lower levels, more and more complexity is transferred and analyzed in the arrows that connect the boxes in the flow diagram, reducing the mystery hidden in the mini mapping theories, that is, the implicit part within the operations. And where does all this end? Recursive decomposition continues towards lower and lower levels until it reaches informational events that are simple and non-further decomposable, called *primitives*.

6.3.2. A Case Study in Visual Perception Demonstrating Recursive Informational Decomposition

Let's return now to our familiar example of visual perception to see how the concepts we have examined are applied when attempting to explain such a phenomenon. We will subsequently examine the case of visually perceiving a horse and how the recursive decomposition of mental operations and the accompanying mapping theories are applied to it. This exemplary case focuses on the operation of object recognition. As shown in *Figure 15* the visual recognition of a horse is a complex operation that occurs in three stages of processing, with the final stage being a combination of perceptual and cognitive factors. We will attempt to focus here on the mental

capacity of object recognition, which is a quite complex mental phenomenon involving many perceptual, mnemonic, and cognitive factors.

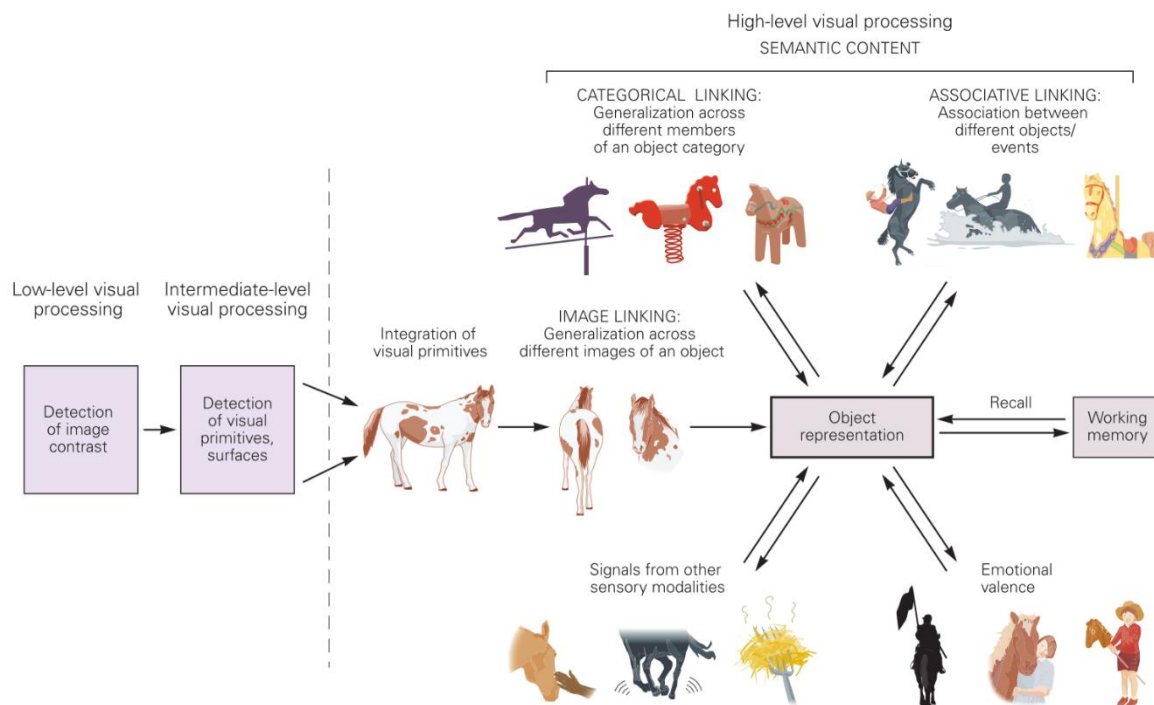


Figure 15: Representation of entire objects in high-level visual processing, reproduced from Kandel, Koester, Mack, & Siegelbaum (2021, p. 565)

The analysis (see Figure 16) begins with the informational description of the (explanandum) mental event (the visual perception of a horse). This description takes the form of a triplet, consisting of an input, a mental operation, and an output. It can be articulated as follows: horse's presentation - operation: perceiving - horse's perception. This description pertains to the upper, undecomposed level (Level A), corresponding to the empirical world containing people, horses, and other empirically observed entities. How does the specific mental operation lead to the perception of a horse? To find out, we decompose it into three sub-operations or processing stages: In the first stage of low-level visual processing, the horse's image contrast is detected. It is followed by a second operation, specifically, intermediate-level visual processing, where visual primitives (that is, local features in a visual scene such as contrast, line orientation, brightness, color, movement, contours, and depth) and surfaces are extracted. Lastly, the observed object is recognized (as being a horse) at the high-level visual processing operation (object recognition). These three operations

occupy the second level, level B, resulting from the initial functional decomposition that took place in level A. If, now, we want to open a little bit more the black box of the mind and shed light on the sub-operations executed in object recognition, we will decompose it and extend the explanation in a third, lower level, level C. Here, things become more complicated, with cognitive and mnemonic operations coming into play. As depicted in the figure, the relations between these operations are replete with feedback loops. Each involved operation both influences and is influenced by the operation of *visually representing* the observed object. Thus, after the initial two operations that feed the object representation box, the *integration* of visual primitives, and the *generalization* across different images of an object (*image linking*), other factors shape and are shaped by the horse's representation: First of all, *categorical linking* generalizes content across different members of the horse's category, such as toy horses, miniature horses, real horses, and the like. At the same time, *associative linking* mediates associations between various objects and events involving horses, such as relevant sports, games involving horses, etc. Both operations serve semantic purposes. Moreover, the representation of the horse is fed by *other sensory modalities* such as hearing, smelling, and touching in a multisensory enrichment operation, and lastly, *emotional valence* is attached to it. The representation is stored in working memory, and at the same time, it is recalled from there for reinforcement or other purposes.

Mental event: *Perceiving a Horse*

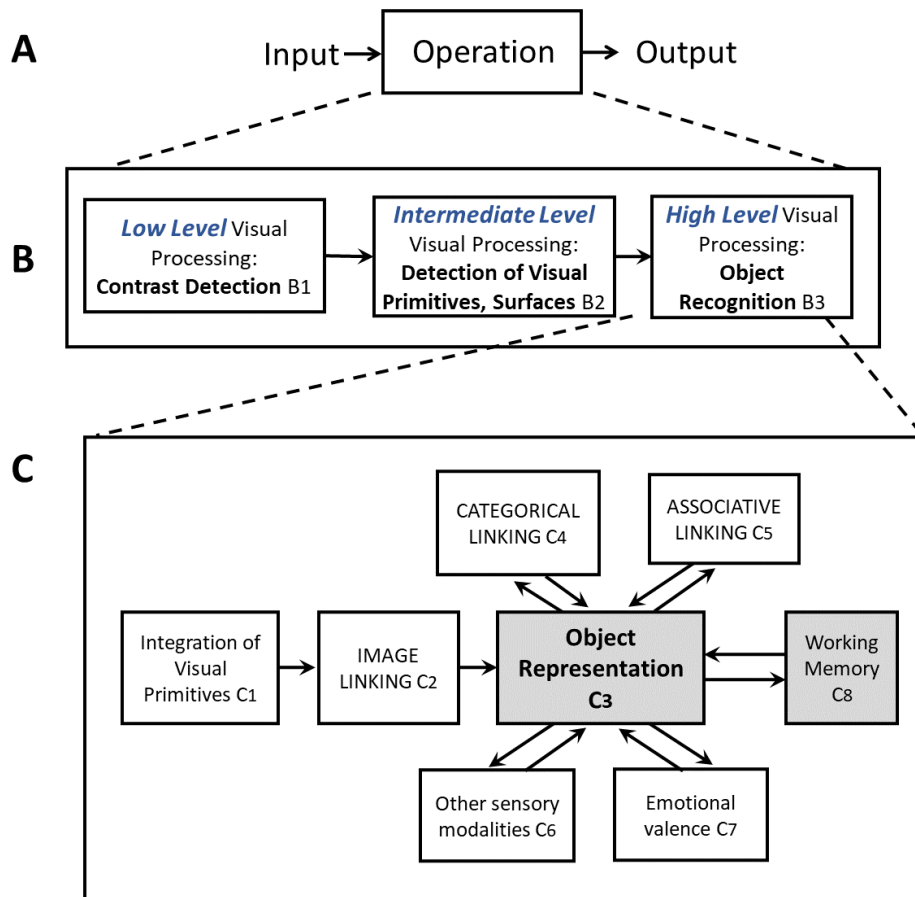


Figure 16: A trilevel decomposition of visual perception, with an emphasis on high-level processing

Putting everything mentioned together, we come up with the following operations:

Level A: Perceiving

Level B: Contrast detection (B₁), visual primitives detection (B₂), object recognition (B₃)

Level C: Visual primitives integration (C₁), image linking (C₂), object representation (C₃), categorical linking (C₄), associative linking (C₅), multisensory enrichment (C₆), attaching emotional valence (C₇), mnemonic operations (C₈)

In the second step of our analysis, we will identify the mapping and mini-mapping theories corresponding to each operation and sub-operation, thus completing each informational event. At the first undecomposed level (A), the mapping theory refers to the horse as the input, which corresponds to whatever is perceived due to its presence, and the output is the perception of a horse. At the second level (B), the mini mapping theories are as follows: the first operation (B₁) receives as input the image reflection that is imprinted on the viewer's retina and outputs an internal imprint of the image's contrast, that is, differences in brightness or color between adjacent regions that help to distinguish objects from their backgrounds. This is taken as input by the second operation (B₂), which outputs a specific arrangement of visual primitives. These are place tokens derived from capturing and processing the image's reflection, like edge segments and edges, bars, blobs, their terminations, virtual lines, boundaries, colored surfaces, and so on (Gordon, 2004)⁷³. All these constitute the input for the third operation (B₃), which outputs the recognition of a horse. At the third level, the operation's C₁ input is a collection of visual primitives, and the output is the specific horse's image. C₂ receives this input from C₁ and outputs to C₃ a generalized image of a horse. C₃ (object representation) has a considerably more complex mini-mapping theory. In addition to the output from C₂ (the generalized image of the horse), it receives as inputs the following: content related to different members (C₄ output) and events (C₅ output) of the horse's category, content related to signals from other sensory modalities (C₆ output), content related to emotional valence (C₇ output) and horse's representations that are stored in short memory (C₈ output). Its output is the visual representation of the horse. The horse's representation feeds as input all related operations, except for C₁ and C₂, which are not involved in feedback loops with C₃. And the remaining operations (C₄ – C₈), since they are connected with feedback loops with C₃, in turn output the informational content that feeds C₃. Everything is summarized in the table below:

⁷³ The hypothesis of visual primitives is based on Marr's computational theory of visual perception and considers how, from an initially primal sketch, the image of the entire object is progressively derived.

Level A Mapping Theory		
Input	→ Operation →	Output
Presentation of a horse	A: perceiving	Perception of a horse
Level B Mini Mapping Theories		
Proximal stimulus	B ₁ : contrast detection	Image contrast
Image contrast	B ₂ : visual primitives detection	Visual primitives
Visual primitives	B ₃ : object recognition	Horse's recognition
Level C Mini Mapping Theories		
Visual primitives	C ₁ : visual primitives integration	Image of <u>this</u> horse
Image of <u>this</u> horse	C ₂ : image linking	Image of <u>a</u> horse
- <u>A</u> horse - Members of the horse category - Events of the horse category - Content from other modalities - Emotions-related content - Stored horse representations	C ₃ : object representation	Visual representation of the horse
Visual representation of the horse	C ₄ : categorical linking	Horse's category members
Visual representation of the horse	C ₅ : associative linking	Horse's category events
Visual representation of the horse	C ₆ : multisensory enrichment	Content from other modalities
Visual representation of the horse	C ₇ : attaching emotional valence	Emotions related content
Visual representation of the horse	C ₈ : mnemonic operations	Stored horse representations

Palmer and Kimchi note that “without some stopping rule, the decomposition assumption could, in principle, be applied recursively ad infinitum” (p. 11), although they consider that in practice, this stops when it reaches a *primitive informational event*⁷⁴ which cannot be further decomposed, but they leave the determination of this point quite vague. Further decomposing the explanandum into even smaller informational events and their corresponding sub-operations and mini-mapping theories proves to be very ambitious and difficult. The reasons for this difficulty are many and relate both to issues concerning the methodology of scientific research in general, and how it has been shaped in fields such as visual perception in recent years. It is generally observed that the evolution of cognitive science into cognitive neuroscience, as Boone and Piccinini (2016) also pointed out, has changed the way cognitive models are constructed, which, especially at the lower levels of description, are increasingly neurobiologically inspired, as one can easily ascertain by simply reading the explanations formulated in modern textbooks of cognitive neuroscience⁷⁵. In recent years, the development of AI-inspired object recognition models has also been significantly involved in addressing the question, as the research on computer vision has exponentially developed and is now used in many online applications, such as Google and Facebook, for self-driving cars, identity detection tools, medical uses, and many other applications. These models exploit entirely new operating principles, such as those of Deep Learning, are in full bloom and now serve in research for human visual perception as reverse engineering methods⁷⁶.

In any case, and on principle, it should be noted that by recursively applying decompositions and as we engage in increasingly detailed investigations, we move further and further away from what can be easily observed and more sophisticated research methodologies are needed to reveal the entities we assume. Everything we assume to explain a cognitive phenomenon are essentially theoretical constructs that will be confirmed when their functional role can be experimentally and observationally linked to detectable physical entities. However, in practice, even this mapping is often quite ambiguous. A characteristic example originating from the field of visual perception relates to the widely accepted hypothesis of the existence of three parallel streams associated with

⁷⁴ As described by Physics

⁷⁵ Consider, for example, the explanatory models of visual perception developed in Kandell, Koester, Mack, & Siegelbaum (2021) and in the chapters explaining sensations and sensory processing in Purves, et al. (2004).

⁷⁶ For a comprehensive review of the latest deep learning-based object recognition architectures and models, please refer to Zaidi, Ansari, Aslam, Kanwal, Asghar, and Lee (2022)

the processing of fundamental features of perceptual content⁷⁷. These streams are as follows: M Stream, responsible for perceptual properties such as luminance, motion, flicker, and hyperacuity; P Stream, concerned with size, color, orientation, texture, and depth; and K Stream, associated with motion, size, color, and texture. In neurobiological terms, these three parallel streams of processing in the Primate Visual System are referred to as magnocellular (M), parvocellular (P), and koniocellular (K). They correspond to cells originating from the retina, passing through specific layers in the lateral geniculate nucleus (LGN) and primary visual cortex (V1), and then projecting to various higher brain areas. These streams converge into two functional pathways that are involved in object localization (where: dorsal stream) and object identity (what: ventral stream). However, the validity of the Parallel Streams Hypothesis, representing the prevalent view in visual perception research, is far from unquestionable. Kaplan (2008) in a rather rigorous and critical stance, highlights several anatomical, physiological, and psychophysical findings that challenge this hypothesis and states that vision scientists often categorize visual properties into concepts like color, size, and motion, and hypothesize that specialized neuronal populations in the brain process these distinct attributes. However, this categorization is a human construct, and the existence of such specific representations in the brain remains hypothetical. He contends that the physical world only consists of space and time, and our understanding of how the brain processes visual information is still largely theoretical. Even if we are unwilling to adopt such a strict stance, we should certainly reflect on whether the entities we invoke as the protagonists of our models correspond to what exists in reality.

Finally, as we have also seen from the example provided, another important factor is that as the repetitions of decompositions increase, the explanatory pyramid grows. Along with this growth, individual questions arise with a geometric progression. How can all these be combined into a coherent, integrated image where findings and conclusions do not contradict each other? In the example we are examining, already at the third level, we have analyzed one of the three operations (high-level visual processing) into 8 sub-operations. If we did the same for the other two, we would likely end up with over 20 sub-operations. If we analyzed those as well, we would reach a three-digit number and so on. This forces us to seriously consider that while recursive decomposition into

⁷⁷ For a thorough review, see Kaplan (Kaplan, The P, M and K streams of the primate visual system: what do they do for vision?, 2008)

psychological entities is theoretically plausible, practically, there are very strong reasons that do not favor it, and after all, it does not even reflect what the scientists themselves do.

Reviewing research in the cognitive neuroscience field, Bechtel (2008) examines how scientific practice has approached two extensively studied cognitive phenomena: memory and visual perception. In these two cases, scientists' approaches diverged significantly. In the first case, they started from the decomposition of individual phenomena that were considered to constitute memory. In contrast, the second case focused primarily on the physical structures that underlie visual perception. Moreover, even within each field and among research teams, approaches varied considerably. Importantly, in both cases, Bechtel's portrayal deviates significantly from a neatly organized and hierarchically structured investigation composed of sequentially decomposed operations, as presented for illustrative purposes in the present section. As he himself notes: "...cognitive investigators seeking to identify operations have generally either remained too close to the phenomena to be explained (drawing upon folk characterizations of mental life), or reached too low (focusing on systems made of neuron-like processing units" (p. 50).

Such diagrams and models are the end product of a lengthy 'battle' between different theories, assumptions, and experimental results. They represent an idealized image that aims to integrate everything we know about a phenomenon into a coherent model, always with the perspective of revising smaller or larger parts of it. There is a continuous dialogue among various methodologies, and each contributes a piece to the puzzle, often rejecting or discarding other pieces from conflicting perspectives. The model described by Palmer and Kimchi is, therefore, a theoretical retrospective interpretation of the structure of cognitive explanations and does not constitute a description of scientific practice. Scientific research operates in a more dynamic manner and certainly not through a top-down preplanned research path. This dynamic nature is evident in the frequent serendipitous discoveries that have shaped scientific progress, the constant interplay between theory and experiment, and the unexpected turns that research often takes. The scientific process is inherently iterative and adaptive, responding to new evidence, technological advancements, and shifting societal needs. Moreover, the collaborative and competitive dynamics within the scientific community often lead research in unforeseen directions, while cross-pollination between disciplines can spark novel insights that were impossible to predict at the outset. Consequently, everything that has been described and explained through examples in these two sections constitutes an idealization, a retrospective representation of the structure of theories within the Information Processing Approach to Cognition.

Building upon this idealized image, I will subsequently explore to what extent the explanatory methodology of the Informational Processing Approach to Cognition can be combined with the constitutive mechanistic account to produce explanations that incorporate physical processes, informational operations, and their contents in an integrated model. To achieve this, we need now to examine the role played by the physical embodiment of an informational processing theory in explaining a mental event.

6.4. TOWARDS “COGNITIVE RICH” MECHANISTIC HIERARCHIES IN EXPLANATORY MODELS

Upon the analysis of the subject, I conducted in this and the previous chapter, I encountered two major challenges that hinder the application of the constitutive mechanistic framework in explaining mental phenomena. The first relates to the inability of the mechanistic explanatory schema to capture the different notions of levels used in explaining cognitive phenomena by cognitive science. The second difficulty pertains to how we can combine the physical with the informational processing and representational aspects of an explanation within an integrated explanatory framework that proceeds with a recursive decomposition of the explanandum, aiming to make explicit part of its complexity and, thereby, partially expose the mental mechanisms concealed inside the black box of the mind. This final section is dedicated to an effort to enrich Craver’s (2007) constitutive mechanistic account with elements from the recursive decomposition of a mental event (as proposed by Palmer & Kimchi, 1986) and from Marr’s (1982) trilevel account. The goal is to address these two challenges.

Constitutive mechanistic explanations aim to explain a phenomenon at an upper level by revealing the mechanisms operating at a lower level that constitute it. Such an explanatory schema typically involves at least two levels: the level where the phenomenon of interest resides and the underlying level that constitutes and explains it through temporally and spatially organized entities and activities. In a purely biological mechanistic explanatory schema, entities correspond to biological parts, and activities are their (relevant to the explanandum) functions. However, when dealing with cognitive phenomena, we must address where cognitive aspects fit within our model of explanation. In information processing systems, these aspects are the description of the mental event as a mapping between an input and an output, which corresponds to its content, and the informational operation performed, which brings about the relevant mental state. Additionally, the third aspect of

a mental event involves its physical embodiment or implementation, in other words, the physical mechanism that instantiates it.

Before delving further into the idea of cognitive mechanistic hierarchies, let's also reconsider Marr's three-level account. As argued (Palmer & Kimchi, 1986; Bechtel, 2008), there appears to be a significant alignment between the three assumptions of the Informational Processing Approach to Cognition that we explore here, namely, mapping theories, operations, and physical embodiment, and the trilevel analysis of an informational processing system by Marr. Specifically, Palmer and Kimchi propose that the mapping theories, which form part of the informational description of a mental (or informational) event, bear a strong resemblance to Marr's computational theories. Furthermore, the Informational Theories resulting from the recursive decomposition of mental/informational events align with Marr's algorithmic level. The correspondence between the two explanatory schemata also holds true for the physical embodiment assumption or the level of implementation (in Marr's terms). What is needed, thereafter, is to attempt an enrichment of the physical mechanism that implements a mental phenomenon with the informational aspects of the phenomenon based on what has already been presented regarding the explanations in information processing systems (Marr, 1982; Palmer & Kimchi, 1986, accounts).

For the Information Processing Approach to Cognition, the physical embodiment assumption (corresponding to the implementation level) concerns the representational and physical vehicles of the mental/informational events that are abstractly described through information mappings and operations. According to Palmer and Kimchi (1986), in IP theories, information is *represented* in the system's states, and the operations that transform it are the physical system's *processes*. They assert that in the physical system that embodies an informational event, representation stands for information, and process stands for operation. In which sense is it considered that a system's state represents an informational state? Palmer and Kimchi mention the most common notion within the Computational Theory of Mind, according to which, representations are being used as surrogates for their referent informational states and preserve their abstract informational structure. The authors do not proceed to an extended and detailed analysis of how exactly the physical implementation is incorporated into the whole explanatory schema, especially considering the difficulties that could arise when recursively applying decompositions; however, this exactly is needed to understand whether cognitive mechanisms are plausible because the central question is this precise relation between the mental and the physical aspect of a phenomenon.

In the final analysis, it appears that we are confronted with the problem of how to relate two separate hierarchies. This is a problem we previously explored when comparing the explanation of motion from a muscle in the elbow joint to the explanation of visual perception: In the case of muscle motion, the explanation should include a hierarchy consisting of only biological entities and processes, while in the case of visual perception, we should account for informational entities and operations, too. Questions that need to be investigated include: Is there a one-to-one correspondence between informational operations and physical processes at each lower level that emerges after their decomposition? In other words, does any explanatory structure produced as a flow chart in an informational decomposition have a counterpart in a structure resulting from physical decomposition? More generally, is there any mapping between physical entities and informational entities that extends all the way down as the phenomenon is decomposed? How exactly are connected the physical and informational aspects of a phenomenon that is explained through recursive decompositions? We will examine two possible solutions.

In the first solution, there is a one-to-one mapping of informational operations and physical processes at each level resulting from the recursive decomposition. Given that the explanatory purposed decomposition of physical processes within the mechanistic explanatory framework results in a mechanistic hierarchy, the term *levels* here means levels of mechanisms. As we have seen, to address the two separate hierarchies problem, that is, how physical implementation connects with a psychological explanation that delves deeper through decomposition, New-Mechanists employed Cummins' (1975; 1983) idea of functional analysis. The notion of functional analysis is sometimes used interchangeably with the analysis of informational operations. However, what functional analysis exactly corresponds to in information processing systems remains somewhat ambiguous. In purely biological phenomena, there is no interpretative issue because functional analysis pertains to biological functions. But in information processing systems, there are two types of functions: biological (physical processes) and mental (informational operations). So, when we use functional analysis in a discussion, which of the two are we referring to? This raises the question: What is ultimately considered a function? Is it an informational operation or a physical process?

In certain perspectives, functional analysis simultaneously serves as both an informational flowchart and an analysis of the functions of physical (biological) parts that implement the informational operation to be explained. That is, it is considered as if these two are the same thing or there exists a strict correspondence between them. I appreciate that this interpretation has its

roots in the elaboration and transformation performed by Craver on Cummins' functional analysis in order to convert it into the model he proposed as norms of mechanistic explanation. He hinted that this could also be applied to cognitive phenomena (e.g., spatial memory formation), although he did not thoroughly analyze how this could be achieved. Prior to this, functional analysis was simply seen as “a view of explanation that is abstracted away from the details of the mechanism that realizes the functions” (Craver, 2007, p. 122). Craver regarded it as an incomplete mechanistic model or mechanism sketch, noting that certain activities and entities, viz, physical processes and entities, are missing, which, when supplemented, would transform it into a complete mechanistic model.

At this point, functional analysis had not yet been explicitly correlated with informational operations. This correlation became evident with Piccinini and Craver (2011), who, utilizing Craver's idea of the mechanism sketch as an incomplete mechanistic model, argued that psychological explanations are actually functional analyses, and functional analyses are sketches of mechanisms. Therefore, they implied that functional analyses concern physical entities since the entities and activities of a mechanism are physical rather than abstract, psychological, or informational entities. Boone and Piccinini (2016), on the other hand, in their two-level distinction, placed cognitive, representational, and computational aspects of a cognitive phenomenon at the functional level, and non-cognitive, mechanistic, and implementational aspects at the structural level. They emphasized that each level of mechanisms is simultaneously structural and functional because it contains structures performing functions. Considering here that in a mechanistic account, structures consist of entities and functions consist of activities, in essence, what they argue is that the functions derived in a functional analysis are the physical activities of a mechanism. All the aforementioned authors imply, therefore, that informational flowcharts and biological functional analyses are identical. In other words, whether we are discussing operations represented in an information flow diagram or spatiotemporally organized activities in a mechanism, we are talking about the same thing; hence, informational operations are simultaneously (or correspond to) physical processes.

If this were true, the physical embodiment of an Informational Processing Theory (in Palmer and Kimchi's terms) or the level of implementation in an information processing system (in Marr's terms) would correspond to the entities constituting the physical mechanism, that is, they would be the answer to the question, 'where does a specific cognitive mechanism take place?' and informational operations would be the mechanism's activities, i.e., the functional roles of the

relevant entities. However, such a proposition cannot hold. Palmer and Kimchi (1986, p. 17) note that “information and operations are abstract, functional entities that exist in the domain of IP descriptions,” thus, they are not physical entities or processes. A close association with physical entities would be overly restrictive. Firstly, if every informational flowchart resulting from recursively applied decomposition represented the spatiotemporally organized activities of the physical mechanism, this would imply that with abstract (i.e., informational processing) methodological tools, we can detect physical processes - an efficient possibility when psychological research aligns with neuroscientific research. When descending levels though, and the informational entities employed are increasingly likely to be fictional, such an association carries the risk of confusing reality and researchers' imagination, as Kaplan (2008) has also emphasized. Furthermore, if this correspondence held, the lower we descend in the hierarchy of informational entities, the more binding our model would become regarding its physical implementation. This is because it would be closely dictated by the informational operations that emerge at each level. Consequently, it would become increasingly challenging to retain the multiple realization principle. Each time, the physical realization would need to preserve the structure of every informational processing flowchart, even those arising at very low levels.

Finally, when we would reach informational events that, according to Palmer and Kimchi's analysis, would be no further decomposable, meaning, they would be *primitive informational events*, we would be forced to consider that their physical counterparts are also non-decomposable, even if theoretically and practically this would appear contradictory. Boone and Piccinini (2016) interestingly suggest that we can have neurocognitive levels all the way down but acknowledge that at some point, we will reach a lower boundary where only neurobiological elements and no longer cognitive ones can be detected. How is this boundary defined? Clearly, a one-to-one correspondence cannot exist between two kinds of entities that simply cease at some point because we have gone too far down for one of them! The dead-end described above has also been emphasized by Lotem Elber-Dorozko and Oron Shagrir⁷⁸, as a fundamental problem in integrating computational⁷⁹ and mechanistic explanations. In their thorough critique concerning how computational and implementational properties can be integrated into the same level of a

⁷⁸ The argumentation of which, I presented in section 6.2.2.

⁷⁹ Referred, here, as well, as “psychological”, or “informational” or “cognitive” explanations

mechanism, they conclude: “It seems that we have here two different hierarchies, one computational, $C_0, C_1, C_2, \dots$, and one implementational, $P_0, P_1, P_2, \dots$ (presumably there are further implementational levels below P_0). But it is left unclear how the two are related to each other” (Lotem & Shagrir, 2018, p. 218). Therefore, this solution is rejected.

In the second solution, instead of directly matching each informational operation to a specific physical process, we divide operations and processes into two levels of mechanisms. Let us assume that the explanandum phenomenon is a mental event of a system (human, animal, robot, or any cognitive agent) that is informationally described through an operation and a mapping theory. I suggest interpreting the physical embodiment assumption by taking it as the implementation of the event by a collection of spatially and temporally organized physical entities and their activities residing at a lower level of organization. This means there is no precise one-to-one correspondence between informational operations and physical processes. The central idea is that instead of identifying which physical process implements an informational operation, we identify which physical mechanism does so. Specifically, we focus on operation-to-physical-mechanism mapping rather than operation-to-physical-process mapping.

Taking into consideration this alternative solution, what would a model combining informational decomposition with a constitutive mechanistic explanation look like? Initially, we start with the undecomposed mental event of a system (e.g., a human). To explain this, we first describe it, mentioning the mapping theory (its content) and the related mental operation. If we want to discover its physical aspect, that is, where and how this specific operation is implemented, we mechanistically decompose it to identify the spatiotemporally organized entities and activities that constitute it, i.e., its physical mechanism. Accordingly, to understand the informational aspect of the mental event, we perform an informational decomposition, revealing the sub-operations and mini-mapping theories that constitute it. That is, we proceed to a psychological explanation of the mental event. Then, to delve deeper, we have two options:

(1) We can continue by further explaining the implementational aspect, describing the mechanisms that bring about each activity. This involves further mechanistic decomposition of every part of our neurobiological mechanism to see how it is constituted by entities and activities at a lower mechanistic level, proceeding purely neurobiologically.

(2) If we are interested in further analyzing the mental phenomenon per se, we perform an informational decomposition to find the sub-operations and their interrelations, along with their

mini-mapping theories, creating an information flowchart at a lower level. Once again, to comprehend how the sub-operations are physically embodied at this lower level, we investigate the set of constituent entities and activities that implement them.

This approach provides a detailed analysis of both the informational and implementational aspects of a phenomenon. Importantly, it does not require a simultaneous and parallel decomposition of both informational and implementational aspects. Instead, in each informational decomposition, we seek anew the mechanisms that can implement the resulting operations.

Let us explore this solution further with an example, referring to the case examined earlier of visually perceiving a horse.

The visual system perceives (mental operation to be explained) a horse (content and part of the mapping theory). This operation is implemented in a set of spatially and temporally organized entities (human eye and early, intermediate, and high-level processing brain areas) performing activities (phototransduction, decoding, and other physical processes) that constitute a mechanism at a lower level of organization. These processes can be further analyzed mechanistically down to subcellular levels or even deeper. Importantly, following the principle of multiple realization, we could explain the visual system of a dog or a robot similarly, all the way down to their respective mechanistic levels, as far as we use implementation details that suit these cases and perform the same (perceiving) operation.

To gain a more detailed understanding of how we see, we conduct an informational decomposition of the mental event that involves a visual system and the perception of a horse. The informational decomposition, as we have already seen in 6.3.2., will reveal that visual perception engages three informational operations: Contrast detection (1), feeding into visual primitives, and surface detection (2), which in turn feeds into object recognition (3). All these are informational operations and have their corresponding mini-mapping theories (listed in the former section).

Now, to connect the psychological to the mechanistic explanation, we need to find how each operation is physically embodied; that is, we seek entities and activities that constitute its mechanism at a lower level. Let us focus on the first informational operation, viz, contrast detection. As derived from relevant theories of visual perception (Kandell, Koester, Mack, & Siegelbaum, 2021), contrast detection is implemented by the eye (entity) converting light to an electrical signal (activity) and early processing brain areas (entity) grouping common features of the visual electrical signal (activity). To further explore this neurobiologically, we start from the first

mentioned entity and activity (eye converting light to an electrical signal) and consider the mechanism that constitutes it, i.e., phototransduction in the eye. This involves entities and activities residing at a lower mechanistic level, such as the cornea focusing light, the iris controlling light that enters the eye, the pupil enabling light entrance, the lens focusing light on the retina, the retina converting light to electrical signals, and the optic nerve carrying the signal to the brain. Again, each of these can be further decomposed to explain its mechanisms in increased detail, and so on. Similarly, we can also mechanistically decompose the other entities and activities that implement contrast detection, namely, early processing brain areas (LGN, superior colliculus, striate cortex), describing their spatiotemporal organization and their activities, too. Thus, we can explain physical mechanisms all the way down, focusing as needed.

But that's probably not all we are asking for. Cognitive phenomena are explored in terms of informational processing, too; in other words, it is very probable that we wish to deepen our understanding of all the sub-operations that bring about the explanandum mental event, too. To do so, we perform another informational decomposition of operations (e.g., object recognition) identified earlier, finding their sub-operations and working similarly⁸⁰.

The proposed explanatory schema exploits both recursive informational decomposition and recursive mechanistic decomposition, allowing the methodological and theoretical interaction between the two different explanatory approaches, viz, psychological and neurobiological theories, and methods, ensuring that at each level, the hierarchies are separated and do not dictate a specific physical realization. Why is such a solution useful in our attempt to combine mechanistic and psychological explanations in a unified mechanistic model? Because it can accommodate levels of analysis (or realization) and levels of mechanisms (or organization, more generally) in one explanatory schema by exploiting both types of decomposition, informational and mechanistic, thus explaining the informational and physical aspects of the explanandum phenomenon all the way down. We will explore all these in what follows. We will begin by examining how the proposed solution aligns with Marr's three levels of analysis. Then, we will delve into how part-whole and realization relations manifest within such an integrated model of explanation.

⁸⁰ A detailed recursive informational decomposition of the case of visually perceiving a horse is presented in section 6.3.2.

This particular solution aligns with Marr's trilevel account, which leaves open how the lower, implementational level of analysis corresponds to the levels of representation and algorithm. Additionally, Marr notes that these three levels are coupled, albeit loosely (Marr, 1982, pp. 24-25):

[...] the different levels at which an information-processing device must be understood before one can be said to have understood it completely. At one extreme, the top level, is the abstract computational theory of the device, in which the performance of the device is characterized as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated. In the center is the choice of representation for the input and output and the algorithm to be used to transform one into the other. And at the other extreme are the details of how the algorithm and representation are realized physically - the detailed computer architecture, so to speak. These three levels are coupled, but only loosely. The choice of an algorithm is influenced for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two.

In the Informational Processing Approach to Cognition, it is stated that: "Mental events can be functionally described as 'informational events,' each of which consists of three parts: the input information (what it starts with), the operation performed on the input (what gets done to the input), and the output information (what it ends up with)" (Palmer & Kimchi, 1986, p. 3). This summarizes Marr's two levels of analysis: the mapping theory corresponds to the computational theory, and the operation corresponds to the algorithmic and representational level. For Marr's third level, addressing *where and how will this algorithm and representation implemented*, a constitutive mechanistic explanation is utilized: a spatiotemporally organized set of entities and activities explains how a specific mental phenomenon (informational operation) with specific content (mapping theory) is executed by a system. In other words, this corresponds to the "detailed computer architecture" that, according to Marr, implements the "abstract computational theory." In terms of a mechanistic hierarchy, Marr's third level, insofar as it explains how the phenomenon physically occurs, is at a lower mechanistic level than the one at which the phenomenon (and its informational content) is described.

Two hierarchies of levels can be detected here. Levels of mechanisms and levels of analysis, where the latter are embedded in the former. Specifically, there are two levels of mechanisms: at the upper

resides the informational event taking place in a system, and at the lower resides the set of its constituent entities and activities. At the same time, three levels of analysis describe the mental event in question: the upper level concerns its content (or mapping theory), which corresponds to Marr's computational theory; the middle level concerns the relevant informational operation, which corresponds to Marr's algorithm and representation level and the lower concerns the physical embodiment or (in Marr's terms) the implementation of the mental event. The two upper levels are embedded within the upper level of mechanisms where the phenomenon taking place in a system is described, and the lower level is embedded in the lower mechanistic level, which stands in a part-whole relation with the upper level and describes the physical mechanism that realizes the mental phenomenon.

Let's extend this idea a little bit more by incorporating Craver's ideas about levels of mechanisms and levels of realization. First, we will see which relations stand between the mechanistic and the realization levels, according to Craver (2007). Subsequently, I will argue that combining Marr's three levels of analysis with Craver's norms of mechanistic explanation does not contradict Craver's initial formulations regarding levels of mechanisms and levels of realization. This holds true as long as we maintain the distinction between realization and composition when theorizing about applying constitutive mechanistic explanation in cognitive phenomena.

Craver (2007) distinguishes between levels of mechanisms and levels of realization, stating that mechanistic levels concern part-whole relations as opposed to realization levels. Drawing from Kim's (1998) account of levels of realization, Craver notes that mechanistic levels are a type of micro-macro levels, standing in part-whole relations. But what about realization, according to Kim's specific account? As Craver mentions, adopting Kim's views, an object S can have two properties, the realized property ψ and the realizing property $\varphi\#$, importantly, both being properties of the same object. Furthermore, we can assume that S has property ψ *in virtue* of having property $\varphi\#$, and for Craver, this means that there is a supervenience relation between ψ and $\varphi\#$ properties. Let us now consider that S is a system and ψ its behavior. The explanandum phenomenon is ψ -ing (by S), and the mechanism that explains it describes a number of entities (Xs) and their activities (φs) that are temporally and spatially organized. In that case, the system has two properties: the property of exhibiting the behavior ψ -ing (property ψ) and the property of consisting of a collection of spatiotemporally organized entities and activities (property $\varphi\#$), in virtue of which it presents the behavior ψ -ing. Property ψ is realized on property $\varphi\#$. This interpretation makes justice to the part-whole relation that Craver claims for levels of mechanisms because the system with its

behavior, namely $S \psi$ -ing, is constituted by the organized collection of activities and entities that correspond to the mechanism that explains this behavior. At the same time, it accommodates the realization relation between the system's behavior and its composition because both are properties of the same system, S . As Craver puts it with an example: "The neuron (S 's) ability to generate action potentials (ψ) is realized by the fact that the neuron (S) has the property ($\phi\#$) of being an organized collection of Na^+ channels, K^+ channels, a membrane, and so on." He continues: "... the ability to generate action potentials is not at a higher level of mechanisms than the property of being a particular organized collection of acting Na^+ channels, K^+ channels, and a membrane. Rather, the parts-plus-organization realize the generation of action potentials" (Craver, 2007, p. 213).

Let's now take another step in the above reasoning and examine whether Marr's trilevel account for analyzing information processing systems aligns with this interpretation that suggests a realization relation within a constitutive mechanism. As we have seen, Marr distinguishes between the computational level, which is about what is computed and why, the algorithmic and representational level, which is about how the specific computational goal is executed and corresponds to the input-output transformation, and lastly, the (hardware) implementation level which concerns how the organized system's physical parts implement the algorithm. How could these levels align with Craver's framework of constitutive mechanistic explanation? Considering what was formulated in the preceding paragraph, the upper (computational and algorithmic/representation) levels correspond to the system's behavior and its content, namely, ψ -ing. This is the property ψ of the system. For example, the visual system perceives a horse. In information processing systems, this corresponds to the input-to-output mapping theory and the relevant operation. The other property of the system that Craver suggested is $\phi\#$ and concerns the organized entities and their activities, which will carry out the explanandum behavior by physically instantiating the necessary processes. $\phi\#$ corresponds to Marr's lower level of implementation.

Summarizing everything presented here, this approach integrates psychological and mechanistic explanations to comprehensively understand mental phenomena. It offers a detailed framework to analyze both the informational aspects (what and how information is processed) and the physical mechanisms (how these processes are implemented biologically), combining Marr's trilevel framework with Craver's constitutive mechanistic explanation. In the specific solution, we achieve a decomposition of the explanandum "all the way down" without encountering the constraints and problems of the first solution. In other words, the psychological model retains its explanatory value

at all levels, and multiple realizability is feasible at each level of the psychological model - not just at the highest undecomposed level tied to empirical reality. Neurobiological research, while inspired by psychology, is not bound by it and can progress independently, subject to subsequent validation of whether neurobiological mechanisms align with informational operations. Furthermore, we are not concerned with the depth of decomposition achievable for each of the two hierarchies, as one decomposition does not depend on the other.

The proposed solution is not unprecedented; as we have seen, Marr's trilevel account has already been suggested as an initial framework for cognitive mechanistic explanations (Zednik, 2017). In the present configuration, the combination with the recursive decomposition of the Information Processing Approach to Cognition, along with compatibility with Craver's insights on levels of realization in mechanistic hierarchies, makes it a promising proposal. However, further exploration is necessary regarding its application to various other cognitive phenomena and alignment with contemporary research across different fields of Cognitive Science. In this dissertation, the aim was to outline this solution as clearly as possible. It is essential to emphasize that such an approach would be convenient only if we fully endorse a computational and modular perspective in cognitive psychology. A critical factor determining the feasibility of explaining a cognitive phenomenon within a mechanistic model is the assumption of decomposability, applicable both to the system implementing the phenomenon and the mental event constituting it. However, as Palmer and Kimchi note, the most controversial assumption of the information-processing approach is that mental events are decomposable. If we do not adopt this assumption, the implications will significantly differ from how they were presented in this investigation.

6.5. SUMMARY AND CONCLUSIONS

In the final chapter of the dissertation, I explored the viability of integrating psychological and mechanistic explanatory models concerning phenomena studied by cognitive science and I gradually centered on the concept of decomposition. I began by reviewing various perspectives on this integration and detected two primary stances regarding the feasibility of fully integrating mental properties into a mechanistic model. One perspective, advocated by scholars like Piccinini, Boone, Craver, and Milkowski, favors the idea of integrating psychological and mechanistic explanations and argues against the autonomy of psychological explanations, suggesting that structural and functional aspects are inherently intertwined. The opposing view, supported by

researchers such as Stinson, Sullivan, Lotem, and Shagrir, acknowledges the close relationship between physical and mental properties but resists reducing explanations solely to mechanistic and implementational aspects. They propose solutions to the dilemma which regard complementary but not seamless integration or they attempt to employ and expand Marr's trilevel account, the latter advocated by Rusanen & Lappi, Zednik, and Bechtel.

In the second section I critically examined these arguments and identified a key issue in approaches that support integration by denying the autonomy of psychological explanations: they often limit mental properties to their functional roles, overlooking the unique perspective of studying the informational and representational aspects of mental events. Psychological explanations so construed, become too narrow to encompass the full complexity of cognitive phenomena. Moreover, I argued that this approach fails to differentiate between cognitive-neuroscientific and purely neuroscientific explanations. Ultimately, denying the autonomy of psychological explanations leads to a dead end. If we take the above into account, it seems that the idea of integration is doomed to fail. However, I also supported the idea of exploring Marr's trilevel account as a potential solution for genuine cognitive mechanistic explanations. Marr's approach, which preserves content at the computational level, analyzes functional roles at the algorithmic level, and describes physical realization at the implementation level, offers a more nuanced perspective on the interplay between cognitive and neural mechanisms.

I continued my exploration in the third section by discussing the concept of recursive decomposition which is part of the Information Processing Approach to Cognition and is used to develop multilevel psychological explanations. Recursive decomposition allows complex mental events to be further decomposed into simpler informational events and their temporal relations. I then presented a case study regarding the visual perception of a horse to demonstrate how recursive decomposition is used in explaining a mental event in multiple levels consisting of operations and sub-operations, from high-level object recognition down to more specific processes. It should be kept in mind, though, that while recursive decomposition is theoretically plausible, practical limitations exist. One such limitation regards the fact that these kinds of models are idealized representations of cognitive explanations, rather than descriptions of actual scientific practice, although very often the scientific research inspired by these models confirms many of the assumptions they contain.

Could recursive decomposition be conceptualized in such a way as to provide a solution for the integration endeavor? In the final part of the thesis, I proposed a solution that builds upon Craver's constitutive mechanistic account and enriches it with elements from Palmer and Kimchi's recursive decomposition of mental events and Marr's trilevel account. I suggested a model that separates informational operations and physical processes into two levels of mechanisms, allowing for recursive decomposition of both aspects without requiring strict one-to-one correspondence. In this approach, the explanandum phenomenon (a mental event) is first described through its mapping theory (content) and related mental operation. To understand its physical aspect, it is mechanistically decomposed to identify the spatiotemporally organized entities and activities that constitute its physical mechanism. For its informational aspect, an informational decomposition is performed, revealing sub-operations and mini-mapping theories. Further exploration can proceed in two ways: (1) continuing with mechanistic decomposition to explain the implementational aspect, or (2) performing another informational decomposition to analyze the mental phenomenon in more detail. This process can be repeated, allowing for a detailed analysis of both informational and implementational aspects without requiring simultaneous decomposition.

I illustrated this approach using the example of visually perceiving a horse, demonstrating how it can accommodate both informational decomposition (e.g., contrast detection, visual primitives detection, object recognition) and mechanistic explanation (e.g., eye structure, brain areas, neurobiological processes). This solution aligns with Marr's trilevel account, with the mapping theory corresponding to the computational level, the operation to the algorithmic and representational level, and the physical mechanism to the implementation level. It also incorporates Craver's ideas about levels of mechanisms and realization, maintaining the distinction between realization and composition.

The framework I proposed, offers several advantages: it allows for decomposition of the explanandum "all the way down" without the constraints of the strict one-to-one correspondence between informational and mechanistic decomposition, preserves the explanatory value of psychological models at all levels, maintains multiple realizability at each level of the psychological model, and allows neurobiological research to progress independently while still being informed by psychology. However, I acknowledge that this approach depends on adopting a computational and modular perspective in cognitive psychology and assuming the decomposability of mental events. If these assumptions are not adopted, the implications would differ significantly from those presented in the present investigation.

CONCLUSIONS

Summary

In my dissertation, I investigated the potential of applying mechanistic explanations to cognitive phenomena. As a foundation for this exploration, I began by delving into the concepts of hierarchy and levels in biological organization. Examining various approaches to defining levels, I highlighted the importance of hierarchical structures in understanding complex biological systems. Next, I explored the concept of decomposition as a tool for analyzing and understanding biological systems. I discussed organization principles related to biological complexity, such as the interplay between aggregativity and emergence, and the role of interactivity and modularity in shaping biological structures.

To demonstrate how these concepts are applied in practice, I delved into compositional explanations in biological systems, using the example of elbow movement as a model for multilevel biological analysis. Through this example, I illustrated how decomposition across multiple levels of organization contributes to a comprehensive explanation of a biological phenomenon. This discussion led to the new-mechanistic approach to explanation, which focuses on identifying and describing the mechanisms underlying an explanandum phenomenon. I offered a detailed examination of the new-mechanistic framework, including its historical roots, key concepts, and the distinction between etiological and constitutive mechanistic explanations.

Having established these foundations, I turned my attention to the specific challenges of applying mechanistic explanations to the cognitive realm. I explored the unique nature of mental phenomena, emphasizing their informational and representational aspects, as well as the multiplicity of levels employed in cognitive science explanations. Through an example of visual perception, I demonstrated how different level notions collaborate in explaining a mental phenomenon. This analysis led me to critically examine the compatibility between mechanistic levels of organization and the various notions of levels used within cognitive science research. I argued that the current mechanistic framework, while effective for explaining purely biological phenomena, faces significant challenges in accommodating the unique characteristics of mental processes. Specifically, it seems difficult for mechanistic levels to capture the representational and

informational character of mental states which is reflected in the involvement of different concepts of levels in their study.

Finally, I explored potential solutions to these challenges, examining various perspectives on integrating psychological and neurobiological explanations. I critically reviewed arguments against the autonomy of psychological explanations, as well as proposals for complementary or unified neurocognitive mechanistic models. I concluded by suggesting a novel framework that combines elements from recursive decomposition in information processing systems, Marr's trilevel account, and Craver's constitutive mechanistic explanation, aiming to create a more comprehensive model for explaining cognitive phenomena. This framework addresses the limitations of the current mechanistic account by separating informational operations and physical processes into two levels, allowing for a more nuanced and multifaceted explanation of mental phenomena.

Conclusions by Chapter

In the dissertation, I presented a comprehensive exploration of the complex interplay between levels of organization, decomposition strategies, and mechanistic explanations in biological systems, with a particular focus on cognitive science. I attempted to systematically build a case for a more nuanced and flexible approach to understanding and explaining complex biological phenomena related to cognition and mental processes.

Chapter One: Here, I laid the foundational groundwork by introducing the concepts of levels and hierarchy in biological systems. I began with an exploration of these concepts through the lens of systems theory and mereology, emphasizing their crucial role in biological inquiry. I reviewed several influential accounts of levels, including the layer-cake model, Churchland and Sejnowski's neurobiological levels, Gibson's ecological levels, and Newell's temporal levels. Through this exploration, it became evident that while these various approaches capture important aspects of biological organization, they fall short of addressing the full complexity of biological phenomena. This limitation underscored the challenge of developing a comprehensive framework for understanding levels in biology that can accommodate the complex and multi-faceted nature of living systems.

For this reason, I delved into Simon and Wimsatt's approaches, which more directly engage with complexity. I presented Wimsatt's ontological account of levels, inspired by Simon's ideas, as a significant step forward in dealing with the complexity of biological systems. It offers a deeper understanding of how various levels of organization interact and influence each other. However,

this account received significant critique from new-mechanistic philosophers regarding its overemphasis on size as a criterion for levels' taxonomy. This criticism highlights an ongoing tension in the field between accounts that prioritize structural features and those that emphasize functional relationships and causal interactions.

I concluded the chapter by introducing the mechanistic levels approach, which defines levels more locally based on compositional relations within specific mechanisms. This approach offers a more flexible and context-sensitive framework for understanding levels in biology, potentially better equipped to capture the dynamic, interconnected nature of biological systems. Through this introductory chapter, I set the stage for the subsequent chapters, which delved deeper into the concepts of decomposition and mechanistic explanation.

Chapter Two: This chapter built upon the foundation laid in the first chapter by exploring the organization of biological structures in greater depth. I focused on the concept of decomposition as a key tool for unpacking the multi-level complexity inherent in living systems. I began by recognizing that biological systems are typical examples of complexity, with many interacting parts across various levels of organization. In this context, decomposition is a versatile strategy that can be applied to both physical structures and abstract functions, allowing researchers to reveal various layers of biological complexity and showing how parts relate to wholes and how structure relates to function.

A central idea I dealt with was the important role of restricted forms of aggregativity in shaping the hierarchical organization of biological systems. I explored how sub-aggregativity plays a crucial role in forming nested hierarchies, allowing higher-level structures and functions to emerge from lower-level components. This principle helps explain how biological systems maintain organization across different scales, from molecular interactions to ecosystem dynamics. I also delved into the interplay between interactivity and modularity, highlighting another crucial aspect of biological organization. I discussed how the strength and nature of interactions between components can lead to the formation of functional modules - semi-independent units within the larger system. This modular organization is a key feature of biological complexity, allowing for specialized functions while maintaining the flexibility needed for adaptation and evolution.

An extensive exploration of the concept of near-decomposability, initially proposed by Herbert Simon, formed a significant part of my argumentation in this chapter. This concept offers a valuable middle ground, allowing for idealized analysis while still capturing both modular and emergent

whole-system properties. I emphasized that even if a system cannot be fully decomposed through functional or structural decomposition, it can still be studied with explanatory tools that presuppose decomposability, as long as it shows characteristics of a nearly-decomposable system. This insight is crucial for the subsequent chapters, as it enables the exploration of how mechanistic explanations can be applied to cognitive phenomena, employing a more flexible decomposability condition for systems that exhibit mental behaviors.

Chapter Three: Here, the focus shifted to the evolution of explanatory models in biological sciences, particularly centering on the development and application of compositional explanations. I began by examining the limitations of traditional nomological and statistical approaches in explaining complex biological phenomena, highlighting the need for a more nuanced and multi-level approach. I introduced Gillett's framework on scientific composition, which provides a robust theoretical foundation for understanding how entities at different levels of biological organization relate to one another in explanatory contexts and how they collectively constitute specific higher-level phenomena.

A significant portion of this chapter was dedicated to a case study on elbow movement, which served as a practical demonstration of how compositional explanations can be applied to complex biological systems. Through this example, I illustrated the process of decomposing a system across multiple levels of organization, from the whole-body level down to molecular interactions. The analysis revealed the importance of selective focus and guided decomposition in constructing effective compositional explanations.

As I navigated through the hierarchical levels, from the broad question of forearm flexion to the specific mechanisms of muscle contraction, the explanation I provided encountered significant bifurcations that shaped the explanatory path. With this, I demonstrated that a comprehensive explanation does not span the entire pyramid by exhausting all possible compositional relations, but rather regards the targeted exploration of explanatory relevant factors across multiple levels of organization. Two critical observations emerged from this analysis: First, the complexity and decomposability of the system varied significantly across different levels, challenging the notion of uniform decomposability. Second, the explanatory power shifted as we moved through the hierarchy, with each level contributing unique and complementary insights. These observations underscored the inherently multilevel nature of compositional explanations in biology, where the explanation emerges from the synthesis of insights across all levels. Thus, the stage was set for the

following chapter, which delved into the new-mechanistic explanatory approach and its application to biological phenomena.

Chapter Four: As the discussion about compositional explanations in biological phenomena continued, I now focused on constitutive explanations structured according to the norms of the new-mechanistic account. I began by tracing the development of mechanistic explanation, from its earlier incarnations in describing and explaining machines to its modern reformulation suited to dynamical and complex systems. I explored the core concepts and definitions that characterize the new-mechanistic approach, centering on Craver's account of constitutive mechanistic explanation in neuroscience.

A significant portion of this chapter was dedicated to exploring the scope of mechanistic explanation by examining various interpretations of the concept and addressing its boundaries and challenges. This included examining what qualifies as a mechanism, addressing borderline cases, and the difficulties posed by highly complex biological systems. I considered the balance between model accuracy and idealization in mechanistic descriptions, extending the discussion of biological complexity from Chapter 2 to inquire how the new-mechanistic approach navigates the tension between decomposability and whole system properties.

I concluded with a detailed case study on spatial memory explanation, demonstrating the practical application of decomposing a phenomenon into a hierarchy of levels, based on Craver's norms for mechanistic explanation. This example illustrated how the new-mechanistic approach integrates the concepts of levels, decomposition, and compositional explanation discussed in previous chapters. It extended across multiple levels of biological organization, from the embodied/situated level where an animal's behavior is observed down to subcellular mechanisms, revealing how constitutive mechanistic explanations can span from observable behaviors to underlying neural processes.

Chapter Five: In this, second part of the dissertation, I initiated a more specialized investigation of the new-mechanistic framework's applicability to cognitive phenomena, with an initial emphasis on the concept of levels. I began by examining the unique nature of mental phenomena and comparing explanations of motor and cognitive behaviors in biological systems. I highlighted intentionality and information processing as the features that distinguish cognitive from other biological phenomena and introduced the vehicle-content distinction, emphasizing that a complete explanatory model needs to account for both the physical substrate and the informational and representational

content of mental states. I then explored how explanations of mental phenomena are provided within the multidisciplinary field of cognitive science. I delved into the multiple perspectives and notions of levels used in cognitive science research, including levels of organization, processing, analysis, description, and realization. Using visual perception as a case study, I demonstrated how different level notions can be combined to explain a phenomenon, utilizing Marr's trilevel account adapted to three stages of processing and three levels of organization.

The main question addressed in this chapter was, whether the mechanistic concept of levels, as proposed mainly by Craver, is compatible with the multiple notions of levels encountered within cognitive science research. My investigation revealed that the model of constitutive mechanistic explanation, in its current form, cannot adequately support cognitive phenomena because the relations that exist between mechanistic levels do not align with those found in cognitive explanations. The challenge of incorporating mental content into levels of mechanisms is identified as a significant obstacle in applying the new-mechanistic framework to cognitive explanations.

I concluded by highlighting the limitations of compositional hierarchy on which mechanistic levels are based, as they proved inadequate in capturing the full spectrum of level notions used in cognitive science. This limitation underscores the need for a more flexible framework that can accommodate various types of level relations beyond mere part-whole compositional relations. I also detected a tension between local and global levels and noted that the local nature of mechanistic levels, while beneficial for specific neurobiological explanations, may be too restrictive for cognitive phenomena that often require consideration of more global, system-wide properties and processes.

Chapter Six: In the final chapter of my dissertation, I explored the viability of integrating psychological and mechanistic explanatory models concerning phenomena studied by cognitive science, gradually centering on the concept of decomposition. I began by reviewing various perspectives on this integration and detected two primary stances regarding the feasibility of fully integrating mental properties into a mechanistic model. One perspective, advocated by scholars like Piccinini, Boone, Craver, and Milkowski, favors the idea of integrating psychological and mechanistic explanations and argues against the autonomy of psychological explanations. The opposing view, supported by researchers such as Stinson, Sullivan, Lotem, and Shagrir, acknowledges the close relationship between physical and mental properties but resists reducing explanations solely to mechanistic and implementational aspects.

I critically examined these arguments, identifying a key issue in approaches that support integration by denying the autonomy of psychological explanations: they often limit mental properties to their functional roles, overlooking the unique perspective of studying the informational and representational aspects of mental events. I argued that this approach fails to differentiate between cognitive-neuroscientific and purely neuroscientific explanations, ultimately leading to a dead end in the integration endeavor. However, I also explored Marr's trilevel account as a potential solution for genuine cognitive mechanistic explanations. Marr's approach, which preserves content at the computational level, analyzes functional roles at the algorithmic level, and describes physical realization at the implementation level, offers a more nuanced perspective on the interplay between cognitive and neural mechanisms.

I then discussed the concept of recursive decomposition, which is part of the Information Processing Approach to Cognition and is used to develop multilevel psychological explanations. A case study regarding the visual perception of a horse was presented to demonstrate how recursive decomposition can be used in explaining a mental event in multiple levels consisting of operations and sub-operations, from high-level object recognition down to more specific processes.

In the final part of the thesis, I proposed a solution that built upon Craver's constitutive mechanistic account and enriched it with elements from Palmer and Kimchi's recursive decomposition of mental events and Marr's trilevel account. In this model I separated informational operations and physical processes into two levels, allowing for the recursive decomposition of both aspects without requiring strict one-to-one correspondence. The proposed approach was illustrated using the example of visually perceiving a horse, demonstrating how it can accommodate both informational decomposition and mechanistic explanation. I concluded the dissertation by highlighting the advantages of this proposed framework: it allows for the decomposing of the explanandum "all the way down" without the constraints of strict one-to-one correspondence between informational and mechanistic decomposition, preserves the explanatory value of psychological models at all levels, maintains multiple realizability at each level of the psychological model, and allows neurobiological research to progress independently while still being informed by psychology.

Throughout all the chapters of my dissertation, I gradually built a comprehensive argument for a more nuanced and flexible approach to understanding and explaining complex biological systems, particularly in cognitive science. I acknowledged the limitations of existing frameworks and

proposed an innovative solution that bridges the gap between mechanistic explanations and the unique challenges posed by mental phenomena. This proposal, though clearly would need further investigation, contributes significantly to the ongoing dialogue in the philosophy of science and cognitive science, offering a path forward in the integration of psychological and mechanistic explanations.

General Conclusions

The central question of the dissertation was the applicability of constitutive mechanistic explanations to cognitive biological phenomena. My investigation focused on two axes: first, on how the mechanistic model of explanation captures the multilevel way of describing, studying, analyzing, and, finally, explaining these phenomena, and how the analytical tool of decomposition functions in their explanation, considering that these phenomena have a dual character: biological and information processing.

This inquiry provided both a negative and a positive account. The negative account addresses the difficulties in combining mechanistic and psychological explanations and is reflected in two specific problems, which I will refer to as the 'levels incompatibility' and the 'recursive decomposition problem'. The positive account proposes how to reconceptualize a cognitive-mechanistic model to achieve integrated explanations without sacrificing the deeper essence of the phenomena being explained and retaining the, valuable for psychological explanations, principle of multiple realizability.

Negative account

Levels incompatibility: The attempt to apply the mechanistic model of explanation, focusing on Craver's (2007) norms of mechanistic explanation and mechanistic levels account, reveals a significant incompatibility concerning the multilevel character of biologically instantiated cognitive phenomena. While this model, in its constitutive form, has been formulated to explain biological phenomena by analyzing them into mechanisms extending to increasingly lower levels, where each lower level of mechanisms constitutes the one above, it fails to fully capture the multilevel character of cognitive phenomena. Specifically, these phenomena, due to their information processing and representational nature, require explanations that employ level concepts beyond the (essentially compositional) mechanistic levels. The informational and representational character of a mental state does not have a compositional relation with its physical substrate.

Instead, they are related through physical implementation, a relation better captured by alternative

level notions, such as Marr's trilevel account for analyzing and explaining information processing systems. Consequently, mechanistic levels cannot be readily applied to incorporate the information processing and representational aspects of cognitive phenomena.

Recursive decomposition problem: The second difficulty pertains to how we can combine the physical with the informational processing and representational aspects of an explanation within an integrated explanatory framework that proceeds with a recursive decomposition of the explanandum, aiming to make explicit part of its complexity and, thereby, partially reveal the mental mechanisms hidden within the mind's black box. The problem that emerged in this investigation relates to how we can combine the two hierarchies of entities that arise from the recursive decomposition of a cognitive phenomenon. These hierarchies pertain to its information processing nature and its physical implementation. If we assume that the decomposition of biological information processing systems results in two hierarchies of entities -one concerning informational entities and the other biological entities- the question becomes: how do these correlate in a multilevel explanatory model? In the case where we consider that the entities of the first hierarchy are directly mapped to those of the second, we are led to a reduction to absurdity. This is because at some point, the informational hierarchy will no longer be decomposable, while the physical hierarchy could continue to be decomposed. It seems rather arbitrary to determine where this one-to-one mapping should cease.

Another problem that arises during the recursive application of decomposition and the attempt to correlate physical and informational entities is that as we descend through levels of decomposition, we may encounter subsystems where the degree of decomposability varies. In some levels, the behavior of the system could be more easily explained in modular terms, while in others it could exhibit whole-system properties due to the complex nature found in many levels of organization in biological phenomena. This problem is linked to the broader concern about the decomposability of biologically instantiated cognitive systems. A possible way to temporarily circumvent this issue would be to consider that even if we find a strongly integrated and complex character in several subsystems, we could still approach them approximately as nearly decomposable.

Positive account

Is it after all feasible to apply the mechanistic explanatory model to cognitive systems, or are we led to a dead end? In this dissertation, I present a proposal for a possible combination of psychological and mechanistic explanatory models. In this approach, while the autonomy and explanatory value

of psychological models are not sacrificed, they could be combined with mechanistic models in a way that provides comprehensive explanations. How can this be achieved? The idea is described as follows.

The framework is grounded in the idea that a cognitive phenomenon can be analyzed through two distinct but interrelated hierarchies: Levels of mechanisms, which focus on the physical implementation, describing the spatiotemporally organized entities and activities that constitute the physical mechanisms underlying the phenomenon, and levels of mental and informational operations, which address the informational aspect, outlining the content of the mental event, its functional role within the cognitive system, and how information is processed. Instead of seeking a rigid one-to-one correspondence between informational operations and physical processes, the proposed framework maps operations to hierarchies of mechanisms. This allows for a more flexible relationship between physical processes and informational operations, acknowledging that multiple physical mechanisms may realize a single informational operation.

My key innovation lies in integrating elements from three prominent accounts: Recursive Decomposition (Palmer & Kimchi, 1986), Marr's (1982) Trilevel Account, and Craver's (2007) Constitutive Mechanistic Explanation. These ensure a hierarchical breakdown of mental events into simpler informational events, the incorporation of computational, algorithmic, and implementational levels, and the grounding of the framework in constitutive relevance. In essence, this "cognitive rich" mechanistic hierarchy employs a two-step approach: Informational decomposition, to understand the informational aspect, where the mental event is decomposed to reveal its sub-operations, mini-mapping theories, and flow of information, and mechanistic decomposition, to understand the physical aspect, where the mental event is assigned to a hierarchy of mechanistic levels that constitute its physical explanation. This approach allows for a detailed analysis of both aspects, informational and physical while maintaining the separation of the two hierarchies. The explanatory value of psychological models is preserved, as each level within the informational hierarchy can be linked to a corresponding physical mechanism or hierarchy of physical mechanisms. Additionally, the framework embraces the multiple realizability principle, recognizing that a single cognitive model can be instantiated by various physical mechanisms.

REFERENCES

- Adami, C. (2009). Biological Complexity and Biochemical Information. In R. A. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 489-511). Springer.
- Adams, F. (2016). Information and cognition. In L. Floridi (Ed.), *The Routledge Handbook of Philosophy of Information-Routledge* (pp. 321-331). London.
- Ahl, V., & Allen, T. F. (1996). *Hierarchy theory: a vision, vocabulary, and epistemology*. Columbia University Press.
- Allen, F. T. (2006). *A summary of the principles of hierarchy theory*. Retrieved 2019, from International Society for the Systems Sciences:
<https://web.archive.org/web/20200430101709/http://www.iss.org/hierarchy.htm>
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450-461.
- Aminoff, M. J., & Daroff, R. B. (2003). Memory, Spatial. In *Encyclopedia of the Neurological Sciences* (pp. 86-90). Academic Press.
- Andersen, H. (2014a). A field guide to mechanisms: Part I. *Philosophy Compass*, 9(4), 274-283.
- Andersen, H. (2014b). A Field Guide to Mechanisms: Part II. *Philosophy Compass*, 9(4), 283-297.
- Ando, A., & Fisher, F. M. (1963). Near-decomposability, partition and aggregation. In A. Ando, & F. Fisher (Eds.), *Essays on the Structure of Social Science Models* (pp. 92-106). Cambridge, Mass: MIT Press.
- Aon, M. A., & Cortassa, S. (2009). Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in. In R. A. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 476 - 489). Springer.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence , & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory Vol 2*. New York: Academic Press.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4), 723-742.

- Bar-Yam, Y. (2019). *Dynamics of complex systems*. CRC Press.
doi:<https://doi.org/10.1201/9780429034961>
- Baysan, U. (2015). Realization Relations in Metaphysics. *Minds and Machines*, 25(3), 247–260.
- Beaney, M. (2021). *Analysis*. (E. N. Zalta, Ed.) Retrieved 2023, from The Stanford Encyclopedia of Philosophy: <https://plato.stanford.edu/archives/sum2021/entries/analysis/>>
- Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Taylor & Francis.
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421-441.
- Bechtel, W., & Richardson, R. C. (1993). *Discovering Complexity Decomposition and Localization as Strategies in Scientific Research*.
- Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press.
- Bertalanffy, L. v. (1969). *General System Theory Foundations, Development, Applications*. New York: George Braziller.
- Bliss, T. V., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of physiology*, 232(2), 331-356.
- Block, N. (2023). *The border between seeing and thinking*. Oxford University Press.
- Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese*, 193, 1509-1534.
- Botvinick, M., & Cohen, J. (1998). Rubber hands ‘feel’ touch that eyes see. *Nature*, 391(6669), 756.
- Brentano, F. (1874). *Psychology From an Empirical Standpoint*. Routledge.
- Bullmore, E., & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5), 336-349.

- Callebaut, W. (2005). The Ubiquity of Modularity. In W. Callebaut, & D. Rasskin-Gutman (Eds.), *Modularity: Understanding the Development and Evolution of Natural Complex Systems* (pp. 3-28). The MIT Press.
- Cat, J. (2017, Fall). *The Unity of Science*. (E. N. Zalta, Ed.) Retrieved 2020, from The Stanford Encyclopedia of Philosophy:
<https://plato.stanford.edu/archives/fall2017/entries/scientific-unity/>
- Celichowski, J., & Krutki, P. (2019). Motor Units and Muscle Receptors. In J. A. Zoladz (Ed.), *Muscle and Exercise Physiology* (pp. 51-91). Academic Press.
- Chemero, A. (2009). *Radical Embodied Cognitive Science*. The MIT Press.
- Chierchia, G. (1999). Linguistics and Language. In R. A. Wilson, & F. C. Keil (Eds.), *MIT Encyclopedia of the Cognitive Sciences* (pp. 92-110). The MIT Press.
- Chirimuuta, M. (2024). *The Brain Abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press.
- Churchland, P., & Sejnowski, T. J. (1992). Neuroscience Overview. In *The Computational Brain* (pp. 17-60). The MIT Press.
- Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *the Journal of Philosophy*, 78(2), 67-90.
- Courtois, P. J. (1977). *Decomposability: queueing and computer system applications*. Academic Press.
- Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.
- Craver, C. F., & Tabery, J. (2019, Summer). *Mechanisms in Science*. (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy:
<https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/>
- Craver, C. F., Glennan, S., & Povich, M. (2021). Constitutive relevance & mutual manipulability revisited. *Synthese*, 199(3-4), 8807-8828.
- Cummins, R. (1975). Functional Analysis. *The Journal of Philosophy*, 72(20), 741-765.

- Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: Bradford/MIT Press.
- Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. C. Keil, & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117-144).
- Dajani, D. R., & Uddin, L. Q. (2015). Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in Neurosciences*, 38(9), 571-578.
- Davies, P., & Gregersen, N. H. (Eds.). (2014). *Information and the nature of reality: From physics to metaphysics*. Cambridge University Press.
- Davis, J. P., Tikunova, S. B., & Janssen, P. M. (2019). Mechanisms of Muscle Contraction and Relaxation. In J. A. Zoladz (Ed.), *Muscle and Exercise Physiology* (pp. 39-48). Academic Press.
- Deacon, T. W. (2014). What is missing from theories of information? In *Information and the nature of reality: From physics to metaphysics* (pp. 146-169). Cambridge University Press.
- Dennett, D. C. (1991). *Consciousness explained*. (P. Weiner, Illustrator). Little, Brown and Co.
- Dretske, F. (1981). *Knowledge and the Flow of Information*.
- Dretske, F. (1983). Précis of Knowledge and the Flow of Information. *Behavioral and Brain Sciences*, 6(1), 55-63.
- Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. MIT press.
- Duda et al. (2019). Human Body Composition and Muscle Mass. In J. A. Zoladz (Ed.), *Muscle and Exercise Physiology* (pp. 3-26). Academic Press.
- Ehrsson, H. H. (2020). Multisensory processes in body ownership. In K. Sathian, & V. S. Ramachandran (Eds.), *Multisensory perception* (pp. 179-200).
- Enoka, R. M., & Duchateau, J. (2019). Muscle Function: Strength, Speed, and Fatigability. In J. A. Zoladz (Ed.), *Muscle and Exercise Physiology* (pp. 129-157). Academic Press.
- Eronen, M. I., & Brooks, D. (2018, Spring). *Levels of Organization in Biology*. (E. N. Zalta, Ed.) Retrieved 2019, from The Stanford Encyclopedia of Philosophy: <https://plato.stanford.edu/archives/spr2018/entries/levels-org-biology/>
- Eysenck & Keane. (2005). *Cognitive psychology: A student's handbook*. Taylor & Francis.

- Floyd, R. T., & Thompson, C. W. (2006). *Manual of Structural Kinesiology*. McGraw-Hill Education.
- Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.
- Fortin, N. (2008). Navigation and Episodic-Like Memory in Mammals. In J. H. Byrne (Ed.), *Learning and memory: a comprehensive reference* (pp. 385-417). Academic Press.
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7), 1293-1326.
- Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. New York, USA: Psychology Press.
- Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis*, 62(4), 316-323.
- Gillett, C. (2016). *Reduction and emergence in science and philosophy*. Cambridge University Press.
- Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44(1), 49-71.
- Glennan, S. (2009). Mechanisms. In H. Beebe, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation*. Oxford University Press.
- Glennan, S., & Illari, P. (2017a). Introduction: Mechanisms and mechanical philosophies. In S. Glennan, & P. Illari (Eds.), *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 1-9). New York: Routledge.
- Glennan, S., & Illari, P. (2017b). Varieties of mechanisms. In *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 91-103). New York: Routledge.
- Gordon, I. E. (2004). *Theories of Visual Perception* (Third Edition ed.). New York, USA: Psychology Press.
- Green, S. (2019, Winter). *Philosophy of Systems and Synthetic Biology*. (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy:
<https://plato.stanford.edu/archives/win2019/entries/systems-synthetic-biology/>
- Guyton, A. C., & Hall, J. E. (2006). *Textbook of Medical Physiology* (11th ed.). Elsevier Saunders.
- Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation.

- Hempel, C. G. (1965). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press.
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of science*, 15(2), 135-175.
- Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4), 500.
- Horgan, T., & Tienson, J. (1996). *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press.
- Hubel, D. H. (1995). *Eye, brain, and vision*. Scientific American Library/Scientific American Books.
- Huxley, A. F., & Simmons, R. M. (1973). Mechanical transients and the origin of muscular force. *Cold*, 37, 669–680.
- Huxley, A. F., & Simmons, R. M. (1971). Proposed mechanism of force generation in striated muscle. *Nature*, 233, 533–538.
- Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, 2(1), 119-135.
- Kaiser, M. I. (2017). The components and boundaries of mechanisms. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 116-130). New York: Routledge.
- Kaiser, M. I., & Krickel, B. (2017). The metaphysics of constitutive mechanistic phenomena. *The British Journal for the Philosophy of Science*.
- Kandell, E. R., Koester, J. D., Mack, S. H., & Siegelbaum, S. (Eds.). (2021). *Principles of Neural Science* (6th ed.). McGraw Hill.
- Kaplan, E. (2008). The P, M and K streams of the primate visual system: what do they do for vision? In R. H. Masland, & T. Albright (Eds.), *The Senses: a comprehensive reference* (Vol. Vision I, pp. 369-381). Elsevier.
- Kauffman, S. A. (1970). Articulation of Parts Explanation in Biology and the Rational Search for Them. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1970*, 257-272.

- Kilarski, W. (2019). Functional Morphology of the Striated Muscle. In J. A. Zoladz (Ed.), *Muscle and Exercise Physiology* (pp. 27-38). Academic Press.
- Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT Press.
- Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11), 826-837.
- Küppers, B.-O. (2014). Information and communication in living matter. In P. Davies, & N. H. Gregersen (Eds.), *Information and the nature of reality: From physics to metaphysics* (pp. 170-184). Cambridge University Press.
- Little, W. (2023). *Introduction to Sociology – 3rd Canadian Edition*. Victoria, BC, Canada: BC Campus Open Publishing.
- Lotem, E.-D., & Shagrir, O. (2018). Computation and levels in the cognitive and neural sciences. In *The Routledge handbook of the computational mind* (pp. 205-222). Routledge.
- Lotem, E.-D., & Shagrir, O. (2021). Integrating computation into the mechanistic hierarchy in the cognitive and neural sciences. *Synthese*, 199(Suppl 1), 43-66.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, 1-25.
- Manson, S. M. (2001). Simplifying complexity: a review of complexity theory. *Geoforum*, 32(3), 405-414.
- Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W.H. Freeman.
- Marr, D., & Poggio, T. (1976). *From Understanding Computation to Understanding Neural Circuitry*. Massachusetts Institute of Technology, Artificial Intelligence Laboratory. A.I. Memo.
- Meyers, R. A. (Ed.). (2009). *Encyclopedia of complexity and systems science*. Springer.
- Miłkowski, M. (2016a). Integrating cognitive (neuro) science using mechanisms. *Pismo Awangardy Filozoficzno-Naukowej*, 2, 45-67.
- Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, 86, 281-297.
- Millikan, R. G. (1991). Perceptual content and Fregean myth. *Mind*, 100(4), 439-459.

- Nagel, E. (1961). Mechanistic Explanation and Organismic Biology in. In *The Structure of Science*. Indianapolis, IN: Hackett.
- Nagel, E. (1977). Functional explanations in biology. *The Journal of Philosophy*, 74(5), 280-301.
- Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
- Núñez, R., Allen, M., Gao, R., Rigoli, C. M., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? *Nature human behaviour*, 3(8), 782-791.
- O'Malley, M. A., & Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1), 58-68.
- Oatis, C. A. (2009). *Kinesiology: the mechanics and pathomechanics of human movement*. Lippincott Williams & Wilkins.
- Oltvai, Z. N., & Barabási, A.-L. (2002). Life's complexity pyramid. *Science*, 298(5594), 763-764.
- Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis.
- Page, S. E. (2009). Understanding complexity. The great courses. *Chantilly: The Teaching Company*. Virginia: The Teaching Company.
- Palmer, S. E., & Kimchi, R. (1986). The information processing approach to cognition. In T. Knapp, & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and Controversies* (pp. 37-77). Lawrence Erlbaum Associates.
- Parry, H. R. (2009). Agent Based Modeling, Large Scale Simulations. In R. A. Meyers (Ed.), *Encyclopedia of complexity and systems science* (pp. 148-160). Springer.
- Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3), 283-311.
- Plato, Phaedrus: 265d - 266a. (n.d.). (B. Jowett, Trans.) Retrieved from classics.mit.edu/Plato/phaedrus.html
- Polger, T. W. (2004). Neural Machinery and Realization. *Philosophy of Science*, 71(5), 997-1006.
- Povich, M., & Craver, C. (2017). Mechanistic levels, reduction, and emergence. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 185-197). Routledge.

- Psillos, S. (2011). The Idea of Mechanism. In P. Illari, & F. Russo (Eds.), *Causality in the Sciences* (pp. 771-788). Oxford: Oxford University Press.
- Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A., McNamara, J., & Williams, S. (Eds.). (2004). *Neuroscience* (Vol. 3). "Sinauer associates." Inc.
- Putnam, H. (1960). Minds and machines. *Minnesota Studies in the Philosophy of Science*, 2, 253-329.
- Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.
- Rall, J. A. (2014). *Mechanism of muscular contraction*. New York: Springer.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551-1555.
- Richardson, R. C. (2001). Functional Decomposition. In R. A. Wilson, & F. C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 329-331). MIT press.
- Robbins, P. (2017, Winter). *Modularity of Mind*. (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy:
<<https://plato.stanford.edu/archives/win2017/entries/modularity-mind/>>
- Roman, F., & Hartmann, S. (2018, Summer). *Models in Science*. (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy:
<https://plato.stanford.edu/archives/sum2018/entries/models-science/>
- Rowlands, M. (2003). *Externalism - Putting Mind and World back together again*. Cambridge, UK, UK: Cambridge University Press.
- Rusanen, A.-M., & Lappi, O. (2007). The limits of mechanistic explanation in neurocognitive sciences. *Proceedings of the European cognitive science conference* (pp. 284-289). London: Francis and Taylor.
- Rusanen, A.-M., & Lappi, O. (2015). On The Scope of Mechanistic Explanation in Cognitive Sciences. *EAPCogSci*.
- Salmon, W. (1971). Statistical Explanation. In W. Salmon (Ed.), *Statistical Explanation and Statistical Relevance* (pp. 29-87). Pittsburgh: University of Pittsburgh Press.

- Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
- Salmon, W. (1994). Causality Without Counterfactuals. *Philosophy of Science*, 61, 297–312.
- Salmon, W. (1999). Scientific Explanation. In M. H. Salmon (Ed.), *Introduction to the Philosophy of Science* (pp. 7-41). Hackett Publishing.
- Schmickl, T., Stefanec, M., & Crailsheim, K. (2016). How a life-like system emerges from a simple particle motion law. *Scientific reports*, 6, 37969.
- Schünke, M., Schulte, E., Schumacher, U., Ross, L. M., & Lamperti, E. D. (2006). *Thieme atlas of anatomy: general anatomy and musculoskeletal system* (Vol. 1). Stuttgart, New York: Thieme.
- Sellars, W. (1963). Philosophy and the scientific image of man. *Science, perception and reality*, 2, 35-78.
- Shadmehr, R., & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental brain research*, 185, 359-381.
- Shoemaker, S. (2001). Realization and mental causation. In C. Gillett, & B. Loewer (Eds.), *Physicalism and its Discontents*. Cambridge : Cambridge University Press.
- Silberstein, M. (2021). Constraints on localization and decomposition as explanatory strategies in the biological sciences 2.0. In *Neural mechanisms* (pp. 363-393).
- Silberstein, M., & Chemero, A. (2011). Dynamics, agency and intentional action. *Humana Mente*, 15, 1-19.
- Silberstein, M., & Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philosophy of Science*, 80(5), 958-970.
- Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467-482.
- Simon, H. A. (1979). Information processing models of cognition. *Annual review of psychology*, 30(1), 363-396.
- Simon, H. A. (1995). Near decomposability and complexity: How a mind resides in a brain. 22, pp. 25-43. ADDISON-WESLEY PUBLISHING CO.

- Simon, H. A. (1998). Discovering Explanations. *Minds and Machines*, 8(1), 7-37.
- Simon, H. A., & Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica*, 29(2), 111-138.
- Smith, J. M. (2014). The concept of information in biology. In P. Davies, & N. H. Gregersen (Eds.), *Information and the nature of reality: From physics to metaphysics* (pp. 123-145). Cambridge University Press .
- Solomon, E. P., Berg, L. R., & Martin, D. W. (2011). *Biology* (9th ed.). Belmont: Brooks/Cole, CA.
- Sporns, O. (2007). Complexity. *Scholarpedia*, 2(10), 1623.
- Sporns, Olaf, and Richard F. Betzel. (2016). Modular brain networks. *Annual review of psychology*, 67(1), 613-640.
- Sterelny, K. (1990). *The representational theory of mind: An introduction*. Basil Blackwell.
- Stern, Y. (2009). Cognitive reserve. *Neuropsychologia*, 47(10), 2015-2028.
- Stinson & Sullivan. (2017). Mechanistic explanation in neuroscience. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 375-388). Routledge.
- Stinson, C. (2016). Mechanisms in psychology: Ripping nature at its seams. *Synthese*, 193, 1585-1614.
- Sweatt, J. D. (2007). Long-term potentiation: A candidate cellular mechanism for information storage in the CNS. In J. Byrne, *Learning and memory: a comprehensive reference* (Vol. 4, pp. 295-326). Academic Press.
- Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11), 5033-5037.
- Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, 96(6), 3257-3262.
- Turrigiano, G. G., & Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5(2), 97-107.

- Varzi, A. (2019, Spring edition). *Mereology*. (E. N. Zalta, Ed.) Retrieved 2019, from The Stanford Encyclopedia of Philosophy:
<https://plato.stanford.edu/archives/spr2019/entries/mereology/>
- Wandell, B. A., & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision research*, 51(7), 718-737.
- Wilson, R. A. (2004). Realization: Metaphysics, Mind, and Science. *Philosophy of Science*, 71(5), 985–996.
- Wimsatt, W. C. (1972). Complexity and organization. *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association*. 1972, pp. 67-86. D. Reidel Publishing.
- Wimsatt, W. C. (1976). Reductionism, levels of organization, and the mind-body problem. In *Consciousness and the brain* (pp. 205-267). Boston, MA: Springer.
- Wimsatt, W. C. (1986). Forms of aggregativity. In *Human nature and natural knowledge* (pp. 259-291). Dordrecht: Springer.
- Wimsatt, W. C. (1994). The ontology of complex systems: levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy*, 24(sup1), 207-274.
- Wu, J. (2013). Hierarchy theory: an overview. In *Linking ecology and ethics for a changing world* (pp. 281-301). Dordrecht: Springer.
- Zednik, C. (2017). Mechanisms in cognitive science. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 389-400). Routledge.