Information-based explanations in cognitive science

A PhD dissertation

by Wojciech Mamak

prepared under the supervision of dr hab. Marcin Miłkowski, prof. IFiS PAN, in the Graduate School for Social Research at the Institute of Philosophy and Sociology, Polish Academy of Sciences

Table of Contents

Acknowledgements **10**
Introduction **13**
Structure of the Dissertation 16
Methodological Approach 19
Chapter 1: Information **22**
Introduction **22**
1. Why Information? Reasons for Informational Optimism 23
2. Counter-IPP 25
2.1 The Siege on the Paradigm 25
2.2 Information Draw 31
3. Desiderata for Explaining with Information Theories 38
4. From Information Theory to Theories of Information 42
4.1 Classical (Shannonian) Theory of Information 43
4.2 Algorithmic Information Theory (AIT) 46
4.3 Semantic Theories 47
4.4 Ecological (Gibsonian) Information 50
4.5 Integrated (IIT) and effective information (EIT) theories 54
4.6 Dretske's Theory 58
4.7 Miłkowski's Correspondence Theory of Semantic Information 64
4.7.1 Infomorphism: The Strict Mapping 66
4.7.2 Infocorrespondence: The Biological Reality 68
4.8 Israel & Perry's Situation-Theoretic Account 69
5. Putting the Desiderata to Work 70
6. Fundamental Properties of Information in Cognitive Systems 74
6.1 Veracity of Information 74
6.2 Objectivity of Information and the Problem of Receiver Relativity 78
6.3 Within Does Not Equal About: A-Information vs. F-Information 82
7. Conclusions 85
Chapter 2: Representation **87**
1. Can't Get Me Out of Your Head (Indispensability of Representations) 87

1.1 Why Representations Anyway?....87
1.2 The Imagery Debate.... 89
1.3 The Representational Job.... 90
1.4 Critique of "Standing-In"....91
2. The Dialectic: Challenges to Representation.... 92
2.1 The Hunger for Representations.... 92
2.2 The Anti-Representationalist Attack.... 94
2.3 The Pluralist Alternative....100
3. The Minimalist Defense: Deflationism....104
3.1 The Argument from Practice.... 105
3.2 Egan's Deflationism....106
3.3 Facchin's Critique of Egan's Deflationism.... 106
4. Information (and/or as?) Representation....107
4.1 Martinez's Theory....108
4.2 Skyrms and the Evolution of Vectorial Content.... 110
4.3 Hiding in Plain Sight. Alistair Isaac on 'Semantics Latent in Shannon Information.... 116
4.4 Rung on a Ladder.... 118
5. Constitutive Properties (The "X" Factors)....121
5.1 Non-Ubiquity.... 122
5.2 Decouplability/detachability.... 125
5.2.1 The Temporal Dimension: Online, Offline, and Detachable....132
5.2.2 The Cognitive State Matrix....132
5.3 Portability.... 136
5.4 Mirroring and Structural Resemblance.... 143
5.5 The Possibility of Misrepresentation....148
6. Conclusion.... 150
Chapter 3: Format....152
1. Why Do We Want or Need Formats?.... 152
1.1 Informational vs. Computational Equivalence....153
1.2 A Short Look Into Engineering. AI & Representation Learning.... 154
2. Philosophical Theories of Format....161
2.1 Historical Accounts....161
2.2 Computational Profiles (Coelho Mollo & Vernazzani)....164
3. When Is Information Explicit?....167

3.1 Procedural Explicitness..... 167
3.2 Easy Access..... 168
3.3. Floors and Ceilings..... 171
4. Empirical Cases: When Format Makes a Difference..... 175
4.1 Tic-tac-toe up to 15.....176
4.2 Cartesian vs. Polar Coordinates..... 177
4.3 Ants on Stilts..... 178
5. Digital, Analog or Generic (Sui Generis) Neural Representations?.....183
5.1 Rulebound Structure Theory..... 183
5.2 Maley's Account..... 184
5.3 Piccinini's Generic Computation.....187
6. Levels of Abstraction: From Interface to Implementation.....190
6.1 Hoffman's Theory.....190
6.2 Martinez's Critique: The Necessity of Integration..... 191
6.3 When Formats Clash.....193
6.4 Practical Concepts.....196
6.5 Pluralism and Redundancy..... 197
6.6 Neural Reuse.....199
7. Implementation & Granularity..... 200
7.1 The Neuron Doctrine..... 202
7.2 Level Up: The Population Doctrine (Hopfieldianism).....203
7.3 Sub-neuronal Accounts..... 206
8. Double Down: The Critique of the Point Neuron..... 207
9. Conclusions - Philosophy of Format Redux..... 208
Chapter 4: Content.....210
1. The Structural Problems.....210
1.1 Directionality of Information and the Arrow of Time.....210
1.2 The Indeterminacy Problem: Neurorabbit..... 214
1.3 The Decoder's Dilemma (Brainjackers and Mindreaders)..... 216
2. The Normative Turn: Fixing Content.....219
2.1 The Failure of Pure Correlation..... 219
2.2 Asymmetric Dependency.....220
3. Teleosemantics (The Backward-Looking Solution)..... 221
3.1 Biological Function..... 221
3.2 Liberal Representationalism..... 222

3.3 The Limits of Tracking..... 223
4. Back to Front: Forward-Looking Theory of Content..... 225
5. The Mechanism of Content: Algorithmic Homuncularism..... 231
5.1 Algorithmic Homuncularism as the Normative Frame..... 231
5.2 The Normative Anchor: Why AH Matters..... 233
5.3 Unmediated Explanatory Information as Content Filter..... 234
5.3.1 Format and Immediacy: Deepening the UE Picture..... 236
5.4 The Empirical Challenge: Algorithmic Coherence..... 237
5.5 Defense: Kinsey's Temporal Decomposition..... 240
5.6 Integrating Situation Semantics: Constraint-Based Content..... 241
6. The Biophysical Implementation: The Two-Point Processor..... 243
6.1 The Mechanical Judge: Solving Semantic Smuggling..... 243
6.1.1 The Two-Point Processor Architecture..... 244
6.1.2 The Biophysical Basis of Error Detection..... 248
6.2 Validation via Coherent Infomax & Precision..... 249
6.2.1 The Computational Objective: Coherent Infomax..... 249
6.2.2 From Mechanisms to Operations: A Computational Taxonomy..... 250
6.2.3 State-Dependent Dynamics: Online vs. Offline Processing..... 253
6.2.4 The Predictive Processing Debate: Error vs. Precision..... 256
6.2.5 Apical Amplification as Precision Weighting..... 257
6.3 Grounding Content: Representations as Rate-Distortion Sweet Spots... 259
6.4 Conclusion: Semantics Without Magic..... 262
Chapter 5: Conclusions of Part 1..... 264
1. The State of the Embattled Paradigm..... 264
2. Partition..... 266
3. The Explanatory Cascade (A Sample Of)..... 267
3.1 Physical Grounding: Minimal Causes..... 268
3.2 Structural Format: Source Coding..... 268
3.3 Semantic Content: Unmediated Explanatory Information..... 269
4. Methodological Pluralism..... 270
5. The Neo-IPP Commitments..... 272
6. Conclusion: The Only Game in Town (Again)..... 275
From Construction to Stress-Testing..... 280
The Logic of the Extension..... 280
Chapter 6: Representing as Compressing..... 283

Introduction: Why compression is imperative....283
1. Regularities, compression, and efficient perceptual coding....284
2. Empirical evidence....298
3. Informational states and representational capacities....303
4. Structural and representational similarities. Infomorphisms?....310
5. Conclusions....317
Chapter 7: Meaning from Matching: How Representational Mechanisms Exploit Structural Similarity....319
1. Introduction: From semantic information to representational geometries....319
2. The Correspondence Network Framework....320
3. Representational Mechanisms Exploiting Correspondence Information....321
4. Unlocking Meaning from Matching: The Role of RSA in Cognitive Science Practice....326
5. Potential objections....341
6. Conclusion....348
Chapter 8: On Generating Information....351
1. Introduction....351
2. Generation vs. faux-generation (fixing the explananda)....352
3. Generativity in neural nets....356
4. Interpolation / Extrapolation boundary....366
5. Computational photography as the method of cases....372
6. Towards the definition....385
7. Potential Objections....390
7.1 The paradox of scale: nominal vs. effective complexity....391
7.2 Biological implausibility....393
7.3 A view from the top: philosophical objections....395
8. Conclusions....398
Final Conclusions....399
1. Philosophy In and Of Cognitive (Neuro)science....399
2. The Stress Tests Recap....400
Representing as Compressing (Chapter 6)....400
Meaning from Matching (Chapter 7)....401
Genuine Generativity (Chapter 8)....401
3. The Explanatory Cascade Revisited....402
4. The Engine of Truth: Verisimilitude and the Consumer....402

4,1 Evading the Empty Symbols (via Consumer Mechanisms).....403
4.2 The Responsible Mindreader: Aligning f-info and a-info..... 403
4.3 The forgotten dimension: operations.....404
5. The explanatory landscape: pluralism, not war..... 404
6. The Fodorian Phoenix.....405
References.....409

Figures

Figure 1: Unequivocal signal transmission (adapted from Dretske, 1981:28)..... 62
Figure 2: Equivocal signal transmission (ibidem).....62
Figure 3: Typewritten characters correspond to Latin alphabet characters (adapted from (Mamak & Miłkowski, 2025))..... 66
Figure 4: An infomorphism obtains between A and B; adapted from (ibidem)..... 68
Figure 5: The Invisible President..... 170
Figure 6: The Visible President; example from Hyvaerinen et al.,2009:4-5).....171
Figure 7: Differences in immediacy based on syntactic and semantic salience of equivalent (or isomorphic) information packages (example from Kulvicki [2010])..... 174
Figure 8: Tic-tac-toe vs. the Fifteen Game..... 176
Figure 9: The schematic of the semantic pointer of TABLE. Multimodal information extracted and processed by the larger neuronal population is gradually compressed to form a more minimal representation. The subsequent evocation of the representation for some other task (e.g., recall) is achieved by the somewhat reversible procedure (compression is lossy, hence the decrease of the resolution'; adapted from Blouw et al.,2015: 1133)..... 288
Figure 10: A diagram from Jóźwik et al., 2016 demonstrating the emergence of a nested hierarchy of semantic categories that mirror the taxonomic order of high-level features..... 333
Figure 11: The RSA pipeline considered from the perspective of information flow.....335
Figure 12: In autoencoders, the encoder part (shown in red) compresses sensory information to compact representations in a latent space. This representation is decoded into sensory data format. The decoder (shown in blue) can be used for counterfactual information generation using a seed chosen from the latent space. The variables z1 and z2 represent the latent variables. [after Kanai et al.,2019]..... 358
Figure 13: Example of image upscaling. The same image of increasing resolution is inferentially upgraded in order to preserve the high fidelity to details even on a bigger scale. The feat is usually not possible using 'direct' image processing.....359
Figure 14: Top-down and bottom-up information transfer..... 361
Figure 15: Yet another example of the same label that could point to separate arrays of information stored at the lower level. Chicken-qua-bird is stored at a different memory address than chicken-qua-meat, even though they are served by the same concept, 'chicken.' Chicken is thus a generative pointer insofar as it could generate one or another, contingent on the task at hand..... 364

Figure 16: Difference between the interpolation (blue) and extrapolation (red) regimes. The former is reserved for inferences between known data points (blue dots) on the regression line, whereas the domain of the latter lies outside of it.....367
Figure 17: The final generated output for all cases of [Albanian Citizen]: A-E..... 380
Figure 18: Cases - D-light (limited distortion); D-hard (more lossy distortion with more features lost); E (almost a seed), respectively, left to right..... 382

Tables

Table 1: The Explanatory Scorecard - table for cross-checking theories of information..... 73
Table 2: The Cognitive State Matrix.....133
Table 3: Taxonomy of Operations in the Two-Point Processor Operation..... 253
Table 4: Tokens, types, and operators inherent in infocorrespondence relations (IMs) within the RSA pipeline..... 337

“Almost all of the material shown here serves to further a view of man as a processor of information, agreeing with my current theoretical disposition. Half of me is ecstatic.” (Newell, 1973:1)

“I like to think of information, at least as a first approximation and what is left from knowledge when you subtract, justification, truth, and belief. It is as if it were an idle thought.” (Dunn, 2008: 581)

Acknowledgements

This dissertation wouldn't exist without a small army of people whose influence and merits in this project are as important as they are diverse.

First, to Zuzia and my family, particularly my Parents—a thousand thank you's for your granite patience and unwavering support throughout this entire ordeal. I cannot express my gratitude enough, but hopefully 400 pages of a text that is of zero interest to you could be some sort of delayed yet delicious dessert.

Then, to my dear friends:
Bartosz, Dominika, and Bazyli—our conversations about philosophy, machine learning, and statistics were more helpful than you probably realized at the time. Certainly much more than *I did* back then.

Andrzej, your illustration of the computer metaphor remains sharper than anything I've found in the academic literature, and that's saying something (nothing).

Łukasz, Jaś, and Emil, you deserve special recognition for never once asking me about philosophy, but always enduring (okay?) earnestly when I suggested we should define what counts as a sandwich. Alhamdulillah to y'all.

From individuals to collectives we go. Thus, the radical evidentialist community, especially the Radewid-3 inner circle, has remained an unrivaled benchmark for philosophy in everyday practice. Similar thanks goes to the various non-philosophical discussion clubs that kept me constantly engaged and definitely argumentative, as the *Scholar's Temper* (pol. *Temperament Badacza; shoutout to all the William Wallace vets*) absolutely demands. Ekipa Mąbiżu, Przygoda w Landzie, Société Des Jacobins, Boryszewska, Mamaghreb and all their outstanding members.

You know who you are, and you know what you did (or you don’t, which is fine too, but you’re being included anyway because you deserve credit).

And, finally, a very special treat-ment must go to Fąfel the Dog, whose contributions, while non-verbal, were indeed informational, contentful, correctly formatted, thus - as we will see - causal and profound for studying the mind in all its forms.

Now to the properly academic cast: Alistair Isaac, Charles Rathkopf, Manolo Martínez, Nick Shea, Johan Hemskerk, Carlos Zednik and Corey Maley all made the journey to Warsaw at various points, enduring the barrage of my questions with remarkable grace. The discussions we had were one of the most profound influences on my thinking about philosophy and cognitive science. The debt I owe to them will be obvious to anyone who even skims through this dissertation.

Lance Nizami, Imran Thobani, but especially Raphaël Millière and Spencer Kinsey provided similarly invaluable help via the internet, proving that intellectual generosity doesn't require physical proximity. Raphaël's input on defining generativity and Spencer's ideas about neural formats were particularly crucial for chapters 8 and 3-4, respectively. Lance even suggested in an e-mail that criticizing the Shannonian mainstream can get you on the CIA's watchlist. That’s why I decided to sell out to Big Information.

The Warsaw academic community, past and current, has been my intellectual home. Particular thanks to Prof. Joanna Rączaszek-Leonardi, Daniel Piecka, Wiktor Rorot, Dr. Paweł Gładziejewski, Tomek Korbak, Prof. Piotr Suffczyński, Ziemowit Sławiński, Prof. Łukasz Dębowski, and all participants of the Philosophy of Cognitive Science Seminar. You've made Warsaw a great fit for this kind of work.

I'm grateful to everyone I met during conferences and study visits, particularly those who sat through early versions of these ideas at Caltech, Antwerp, Sorbonne, Turin (big thanks to Nico Orlandi here), Cracow, Sofia, Ulaanbaatar, and others. Special recognition goes to the organizers of Neural Mechanisms Online and the School of Ideas in Neuroscience—these remain the benchmark events for me at the philosophy/neuroscience intersection.

Lecturers and participants I had the pleasure of interacting with there (in random order): Max Collard, Kacper Kondrakiewicz, John Krakauer, Adrienne Fairhall, Michał Wierzchoń, Pamela Lyon, Carina Curto, Philippe Haueis, Kate Nave, and Bryce Counts - they all influenced my thinking tremendously and refueled my passion for this research when I needed it most.

Hamza Giaffar, Nedah Nemati, Amira Moeding, and Mira Marcinów deserve special mention—they now occupy that rare category of people who are both friends and profound intellectual influences.

Thanks also to the anonymous reviewers who read parts of this dissertation and improved it with their feedback.

Finally, and most importantly, Marcin Miłkowski, my supervisor, without whom literally none of this could have started, continued, or finished. Our relationship predates even the painfully long and arduous creation of this philosophical pyramid (scheme). He got me into this type of research during my second year of studies with a (seemingly harmless, but it turns out, fateful) lecture on anticipatory representations, which, as it turns out, still haunt these pages (Section 6.2.2 of Chapter 4). His basalt (or whatever trumps granite) patience, crushing but fair and super constructive criticism, and titanic effort have shaped not just this dissertation but the entire trajectory of my thinking. The debt I owe him is beyond what these acknowledgments can express.

Introduction

This project examines the explanatory import of information and related concepts for understanding cognitive systems in modern mind and brain sciences. It is a piece of philosophical work that falls under the philosophy of cognitive science, but it is also an attempt at philosophy *in* cognitive science—a distinction I owe to Danielle Williams (see: Brooks, 2009). I firmly believe that philosophical tools and concepts should not only analyze scientific practice ex post, but can actively contribute to understanding cognitive systems, particularly at the intersection of philosophy and neuroscience.

Information processing stands as a rare consensus point in cognitive science: most theorists agree that cognitive systems process information in some sense. Cognitive (neuro)science famously lacks unification and conceptual consistency, remaining in search of an overarching theory. It is a Babel tower of disciplines, vocabularies, tools, and concepts. Yet if we were to identify shared truths across this fragmented field—bets on statements that most theorists would agree upon regardless of their location within it—claims like "cognitive systems are in the business of information processing" or "the basic building block of cognitive agents is information" would be prime candidates for rare uncontroversial common ground.

So perhaps it is peculiar to provide a philosophical analysis of this part of the Babel tower. Why not enjoy a rare victory? Why kick the tires on a sturdy (sic!) vehicle? Well, partly because the status of information processing's role in cognition is not that simple on closer inspection. This apparent agreement masks profound rifts, underspecification, and illusory coherence based on fast-and-loose playing with the mighty general word, ending with a massive, sprawling, equivocation. The theoretical landscape reveals vast disputes about information's ontological status, the scope of information processing explanations, and the role informational

concepts should play in explaining cognition. Some theorists claim information processing is involved in virtually all cognitive activity, while others restrict it to "higher order" tasks, distinguishing it from action and motor control. Famous delineations like System 1/System 2 processing (Stanovich & West, 2000; Sloman, 2011; Evans & Over, 2013; for critical discussion, see Osman, 2004;2018; Keren & Schul, 2009), dual-coding theories (Paivio, 1986), and pluralist but computational views that include both ecological and representational modules (Orlandi, 2014) attempt to accommodate different regimes of processing within cognitive architectures. Still others argue that information talk is epiphenomenal—useful for pragmatic purposes but lacking genuine causal or biological significance. There is no consensus on how to adjudicate the relationship between information processing and dynamical systems theory (DST), perceptual control theory (PCT), or Gibsonianism in its novel versions. Some think they are competing and irreconcilable frameworks (van Gelder, 1995; Thelen & Smith, 1994), whereas others claim they are different modes of description (Bechtel, 1998; Eliasmith, 1996), could be unified in principle (Mediano et al. 2021; Balduzzi & Tononi, 2008), or conjoined to provide pluralistic explanations (Dale, 2008; Dale et al., 2009). Radical enactivists (Hutto & Myin, 2013; Chemero, 2000, 2009), some proponents of predictive processing, and certain Gibsonians reject the need for information processing altogether, viewing it as a misleading and outdated framework that should be discarded.

I will defend the thesis that there is an explanatory strategy in cognitive science that I call the 'Information Processing Paradigm' (IPP) (or Information Processing Theory; used interchangeably) and that it remains the most robust, promising, and empirically grounded approach to understanding mind and brain. I track the historical reasons for its creation, comment on its motivations and merits, but primarily analyze its commitments, implicit and explicit assumptions, and assess

critiques from behaviorism, ecological perception, dynamicism, interactivism, predictive processing, and 4E cognition. Furthermore, I do not claim that original formulations of IPP were faultless or that the paradigm needs no improvements. On the contrary, my main aim is to lay out how to be a good information-flow cognitive scientist[1] after these various theoretical challenges—to spell out and defend a neo-IPP, an expanded and improved version of the Marlboro man of cognitive explanations, which never should (and indeed does not) become unfashionable.

This project intentionally brackets several related debates to maintain focus on informational explanation in cognitive neuroscience. I set aside questions of conscious versus unconscious processing (including integrated information theory), deeper treatment of computationalism (such as the problem of indeterminacy of computation; Fresco, 2021), physical symbol systems, the rich literature on genetic information in philosophy of biology (especially e.g., Griffiths, 2001, which has direct repercussions for the explanatory status of information in biology generally; Godfrey-Smith & Sterelny, 2012), debates on the applicability of informational concepts in animal cognition (Kalkman, 2019; Mann, 2018; Krstić, 2025), modularity and cognitive penetrability (particularly tempting whenever Jerry Fodor appears via his informational encapsulation ideas), and general philosophy of science debates on the nature of explanation. I draw substantially from literature on computational explanations, causal accounts, levels of abstraction, and

---

[1] I directly hint to a passage from Fodor (1975: 74, footnote 15) where - in a rare move! - explicitly labels an IPP proponent: A—by now—chestnut of a question that is supposed to embarrass **information flow psychologists** [bolding is mine] goes like this: 'If you are willing to attribute regularities in the behavior of organisms to rules that they unconsciously follow, why don't you say (e.g.) that the planets 'follow' Kepler's laws in pursuit of their orbits about the sun?' The point, of course, is to suggest that the only real case of rule following is conscious rule following by articulate organisms. What other organisms do is (not follow rules but) merely act in accordance with them". Sidestepping the rationale of the quote (which will come back big for the IPP), seems to me like a good coinage for an honest-to-God IPP proponent. Interestingly enough, information talk and information theory authority were so prevalent for the majority of the GOFAI (Good Old-Fashioned Artificial Intelligence) era of cognitive science that there did not seem to be a need for such a label.

neomechanism, but remain far from comprehensive treatment of any of these. These choices are intentional rather than oversights—engaging these topics comprehensively would dilute the central analysis, which is especially dangerous for a project operating at this scale while also being a meta-level philosophical account of actual brain science.

I have also resisted expanding on certain themes that represent unfinished ends deserving future pursuit. The connection between information processing and inferentialism is a hugely underexplored topic that needs further focus (but see Wansing & Shramko, 2019 and recently Laukaityte & Colombo, 2025). Similarly, the relationship between informational concepts and debates on the format of mental representation, while touched upon here, opens avenues for more in-depth investigation that extend beyond this project's scope.

Structure of the Dissertation

This dissertation follows an unconventional two-part structure that I liken to an Olympic ice-skating program—a metaphor suggested by my supervisor that I took not merely as an ad-hoc rationalization but as a guiding principle, a metaphor to live by (Lakoff & Johnson, 1980). It causally influenced the design, write-up, and construction of this thesis. I followed this rule, not merely acted in accordance with it—a distinction that, as we will see later, is crucial for understanding informational structures.

In competitive figure skating, performances consist of two parts: a short program of required technical elements demonstrating mastery of fundamentals and a longer free program where contestants take greater creative risks while building on those fundamentals. The former is geared toward demonstrating mastery of the rudiments of the discipline, allowing for limited and specific merit: thoroughness in

executing known features and some creativity in combining the already-known for a fresh take. Then comes the long, free part, where the contestant is given a much longer leash and encouraged to take the basics forward, providing novel contributions and experiments. The constraints for this part are looser and judged according to technical soundness, but much more interest is directed toward novelty and creative expression. Similarly, philosophical dissertations in the Polish tradition should demonstrate both thorough command of existing debates and novel contributions that advance the field.

Hence, the reader will find the entire thesis firmly divided into Part 1 and Part 2. **Part 1** serves as my compulsory program in the philosophy of information, especially in the context of cognitive science and its explanatory strategies. I discuss several main theories of information in detail, examine the motivations and ramifications in the philosophy of information (with minor historical context), and analyze the relationship between information and other fundamental explanatory concepts: content, representation, and format. While this part may at times seem encyclopedic, it contains several contributions. The discussion of format's role in cognitive information processing is largely novel and frequently ignored in the literature. My introduction of f-information (information-for-the-system, sometimes called information-within) and a-information (information-about-the-system) is original, though it mirrors some discussions of external/internal information and information as substrate versus measurement—debates usually treated as secondary or tacitly implied. I spell this distinction out clearly and discuss its ramifications in detail. This distinction has deep connections to the neural decoding research program that implements a sort of responsible mind-reading—a goal toward which this thesis aims to bring IPP.

Most importantly, I strongly commit to connecting philosophical discussion with empirical neuroscience research, focusing on neurally instantiated brain

information. This represents what I believe to be a merit of this thesis, and I follow here the approach exemplified by thinkers like Cameron Buckner, Gualtiero Piccinini, and the philosophical community centered around Neural Mechanisms conferences and Brains Blog. This engagement with actual brain science provides crucial context for assessing recently reignited debates on encodingism and neural coding, which hinge on a solid understanding of what kind of information we have in mind and how well specified the concept is.

Furthermore. I believe that there is genuine merit in clarificatory work, especially regarding concepts as troubled by underspecification as information. Conceptual clarity does not end with careful analysis of the notion of information itself, but also benefits greatly from investigating its links with other fundamental explanatory concepts in cognitive science. The comprehensive analysis of information's relationship with content, representation, and format should help us understand the conditions required for information-based explanations in cognitive science to succeed. To foreshadow the conclusion, I believe the informational picture is still the only game in town (to paraphrase Fodor). Part 1 advocates for this position while acknowledging genuine opponents.

Part 2 surely corresponds to Part 1—certain strands, ideas, and concepts return (such as infomorphisms, compression, and the Taking condition)—but should be treated as virtually independent. This part serves as my free program: three separate chapters that advance the state of the art in philosophy of cognitive science and philosophy of information. These chapters tackle limited but vitally important problems that have received surprisingly scant attention, despite being crucial for assessing the explanatory import of informational concepts for cognitive systems. Chapter 7 has already been published in a broader version (in cooperation with Prof. M. Miłkowski), while chapters 6 and 8 are penciled for submission. They

represent specific, original contributions to the field, taking forward the consensus view established in Part 1.

Methodological Approach

I must acknowledge that philosophical treatment of something so vast limits the ability to provide highly targeted, ultimate answers, which may render the enterprise less satisfying than readers with finer-grained preferences might appreciate. I can concede that and perhaps write it off as an aesthetic preference. There is a well-received distinction in cognitive science history that we can borrow and repurpose here: theorists fall roughly into two categories—scruffies and neats. First suggested by Roger Schank in the mid-1970s and widely applied within the early cognitive science research community, this distinction delineates rigid, rule-based, highly regimented approaches from more freewheeling, big-picture methodologies. A complex, multifaceted problem such as assessing the applicability of informational concepts to cognition broadly construed strongly pushes one toward being "scruffy." I hope that even readers who locate themselves closer to the "neats" end of the spectrum will find the arguments sound and rigorous, even if they feel less structured than expected (which would be especially ironic when advocating for structural representations).

The literature in this area is remarkably disjointed, operating across multiple levels (from sub-neuronal to complex networks) and happening across many fields—it can hardly be called a unified battlefield. It is more a patchwork that I have strived to sew together, looking for essential links and ways to reorganize the debate on the explanatory status of information for cognitive systems. If I have succeeded, the reader will find not only a defense of IPP but also a clearer map of the theoretical

terrain and the genuine choices facing cognitive science as it continues to grapple with its most fundamental concepts. I sincerely hope the Reader will find this structure acceptable and perhaps even illuminating.

To wrap up the rundown of the structure, the order of sections is as follows:

Part 1: The Foundations (Theoretical Grounding) establishes the conceptual pillars of the Neo-Information Processing Paradigm.

Chapter 1 defines the concept of information, moving from Shannon to semantic and structural accounts. Chapter 2 defends representation, identifying its constitutive properties (decouplability, structural mirroring) against eliminativist critiques. Chapter 3 examines format, arguing for the computational necessity of specific encoding structures. Chapter 4 defines content, defending a forward-looking, normative account of semantic determination. Chapter 5 synthesizes these arguments into the "Neo-IPP" framework.

Part 2: The Stress Tests (Empirical Validation) applies this framework to concrete empirical domains.

Chapter 6 investigates compression as the engine of representation. Chapter 7 explores meaning from matching through Representational Similarity Analysis (RSA). Chapter 8 analyzes generativity, defining the conditions for genuine informational novelty in computational systems.

PART 1: THE INFORMATION PROCESSING PARADIGM (IPP)

Chapter 1: Information

FROM INFORMATION THEORY TO THEORIES OF INFORMATION

Introduction

The aim of this chapter is twofold. Firstly, I intend to revisit the debates within philosophy and cognitive science on the nature of mental information. Secondly, by doing that, I intend to set up the current philosophical theory of information for evaluation of how it fares when applied in the neural domain in the next chapter. I open with a brief discussion of the origins and reasons for enthusiasm about the Information Processing Thesis, understood as a conviction that minds are primarily explainable through information processing. Historically, proponents of this view—whom Jerry Fodor (1975, p. 74) aptly dubbed "information flow psychologists"—have argued that intelligent behavior cannot be explained merely by the organism acting in accordance with laws, but requires the organism to *follow* internal rules represented within the system.

Then, I review classical (Shannonian) theory (Shannon, 1947, Shannon & Weaver, 1949, Shannon, 1963), its algorithmic extensions (Kolmogorov-Solomonoff-Chaitin; Chaitin, 1979; sometimes conceptualized as Minimum Description Length, Grünwald & Roos, 2019) semantic alternatives (Bar-Hillel & Carnap, 1953, Hintikka, 1970), ecological (Gibsonian) information (Gibson, 1966, 1979/1986), Dretske's account (Dretske, 1981, 1983), and Miłkowski's (2023) structural correspondence theory. Later

I discuss the question of veracity (truth-evaluability) and objectivity of informational states in the context of the so-called receiver-relativity problem (Rathkopf, 2017, Rathkopf, 2018). On this basis I criticize the 'hydraulic metaphor' of reified information flow. Subsequently the question of whether a complete theory of information should be all-encompassing or rather taxonomical is addressed. As a result, the ground is cleared for the follow-up analysis of the relation between information processing, mental content, and representation in the neuroscientific context through later parts of the thesis. Through integrating results from both philosophical analyses and empirical studies, I will be particularly interested in the problem of whether identifying information flow can lead us to explaining the representational powers of a given cognitive system. And if not, what is the information-centered explanation lacking?

1. Why Information? Reasons for Informational Optimism

Firstly, it is important to address the question of why we should be committed to informational analysis of cognition in the first place. That may strike us as an odd question. The view of cognitive systems as essential information processors (as the necessary condition of cognition) is probably as mainstream of a view as it gets in such a heterogeneous field like cognitive science. So, is it really theoretically interesting to analyze this supposedly widely accepted assumption? I am strongly convinced that the answer is 'yes.' And it is precisely because of its omnipresence that the notion of information in cognitive science became dangerously vague at best and ambiguous at worst, purporting to perform a lot of explanatory work as a free-for-all, not well-defined nor critically understood. That poses a significant problem, as information talk bears a lot of the burden of explanation in core subdisciplines of cognitive science, such as psychology and neuroscience, while the

semantic field of the term is not well mapped. That calls for thorough philosophical analysis.

To clarify the target of this analysis, we must move beyond vague usage and define the paradigm formally. The Information Processing Paradigm (IPP) is not merely a loose metaphor; it is a rigorous research tradition defined by two necessary and sufficient conditions. First, it accepts the foundational commitment that cognitive processes are *constituted* by the encoding, storage, retrieval, transformation, or transmission of information. Second, it adheres to the architectural commitment that this processing occurs via a "stage model,"[2] where information flows sequentially through functionally discrete components. As Mulder articulates, "According to the information processing approach these processes occur in a sequence of ordered stages... A flow chart is a sequence of boxes where each box reflects a stage of processing" (Mulder, 1983: 19)

However, our commitment to this framework should be tempered with epistemic humility. As Dominic Massaro (1990) famously remarked, "Analogous to Winston Churchill's view of democracy, information processing is not perfect but it is the best that is currently available" (p. 135). Accordingly, IPP has not been left unscrutinized (and indeed if it were, it would probably not merit a substantive defense, as we mount in this thesis). Let us then briefly review the critics of the IPP

---

[2] While often retrospectively labeled the "stage model" or "stage theory" in textbook expositions (e.g., Standring, 2025; Huitt, 2003), the original authors used distinct terminology: Broadbent (1958) referred to his framework as "filter theory", while Atkinson and Shiffrin (1968) described a "system" comprising "structural features" (sensory, short-term, and long-term stores) and "control processes". The "stage" label remains accurate, however, because these models explicitly operationalized cognition as a linear sequence of discrete processing steps—visualized famously by Broadbent's use of engineering flowcharts. This conceptual architecture established the logic of the "computer metaphor" that persists in modern explanatory strategies like Nick Shea's "Algorithmic Homuncularism". Just as the stage model decomposed memory into discrete temporal steps, Algorithmic Homuncularism requires that neural function be explainable via spatial and temporal decomposability—the commitment that specific cognitive sub-routines can be mapped onto spatially distinct neural regions operating over distinct timescales. We will draw from Shea's strategy extensively throughout the dissertation.

(broadly construed) in order to first, better understand the deficiencies of the framework and get ready to ameliorate it.

2. Counter-IPP

Indeed, the dominance of the Information Processing Paradigm (IPP) in cognitive science has long functioned as a unifying framework (from its conception, really), offering a computational language to bridge the gap between the wetware of the brain and the software of the mind. However, a convergent body of recent scholarship has launched a critical interrogation of this paradigm, arguing that it rests on a category error: the reification of engineering metaphors into biological realities.

2.1 The Siege on the Paradigm

Despite its dominance, the IPP has faced a sustained barrage of criticisms from theorists who argue that the computer metaphor fundamentally misrepresents biological reality. This critique operates on three levels: the nature of memory, the nature of the organism-environment coupling, and the nature of the neural signal itself.

At the macro level, Robert Epstein (2016) argues that the IPP is a linguistic trap that blinds us to biological reality. In his essay "The Empty Brain," he rejects the "storehouse" view of memory implicit in the stage model. He writes, "We don’t store words or the rules that tell us how to manipulate them... and then transfer the representation into a long-term memory device". Using the example of drawing a dollar bill from memory, Epstein demonstrates that students cannot reproduce the bill accurately despite seeing it thousands of times. This proves, he argues, that

there is no "file" stored in the brain to be retrieved; rather, the organism is changed by experience and engages in a dynamic, imperfect reconstruction. For Epstein (2016), the IPP is "shoddy thinking" that obscures the non-symbolic reality of the organism.

This skepticism also reverberates in the form of Radical Embodied Cognition (Wilson & Golonka, 2013). They critique the standard IPP assumption of the "poverty of stimulus"—the idea that the brain is "locked away inside our heads with only impoverished... perceptual access to the world." Because IPP assumes input is poor, it posits elaborate internal mental representations to enrich it (the so-called "cognitive sandwich"). Wilson and Golonka reject this, proposing an alternative framework. According to them, "embodiment is the surprisingly radical hypothesis that the brain is not the sole cognitive resource... our bodies and their perceptually guided motions... replacing the need for complex internal mental representations".

This critique draws heavily on H.H. Pattee's (1974) distinction between specificational and indicational information. Critics argue that the environment often *specifies* appropriate action (constraining the organism physically/dynamically), whereas the IPP (the story goes) wrongly assumes the environment only *indicates* action (requiring the organism to compute a decision). If the environment specifies the action—as in the case of a baseball player catching a fly ball using optical trajectories rather than internal physics calculations—the internal processing stages posited by the IPP are superfluous.

Yet another strategy is undertaken by Mark Bickhard. In a series of articles (2000, 2009a, 2015, 2017; among many) and books (Bickhard & Terveen, 1995;), he attacks the "what" of IPP by exposing the logical incoherence of "encodingism" (Bickhard, 2009b). He claims that the IPP (note here that he does not use the term directly) assumes that mental representations "stand in" for environmental features.

Bickhard argues this is circular: to establish that internal state X encodes external state Y, the system must already know Y to check the code. This empty symbol problem leads to an infinite regress of interpreters. Instead of static encoding, Bickhard proposes an "interactivist" model where representation emerges from the anticipation of interaction outcomes—a functional, forward-looking relation rather than a backward-looking causal correspondence. Interestingly enough, he foreshadows ideas put forward recently by Buckner[3] (2022).

This ‘reification’ line of critique was recently revived and intensified by Romain Brette in an influential paper, which dissects the "coding metaphor" as a barrier to causal understanding (Brette, 2019). Brette contends that "neural coding" is not a neutral scientific term but a metaphor that smuggles in a fatal logical requirement: the presence of a decoder. A code, by definition, is a mapping between a signal and a content. For a neural spike train to function as a code for the organism (rather than just for the scientist), there must be a downstream structure capable of "reading" that code and recovering the original content. However, Brette observes that "a neural code is information only by reference to something with a known meaning," which fatally implies the existence of a homunculus—an internal observer who knows the code. Because there is no such observer in the brain, the "code" exists only in the mind of the scientist. Brette proposes abandoning the view of spikes as "hieroglyphs" to be deciphered in favor of viewing them as actions that perform physical work on downstream targets, driving the system through a circular, dynamic loop rather than a linear channel. Since there is no downstream "little man," he argues that the current usage of "neural code" is a categorical error that mixes causal and representational language without justification.

---

[3] Although Buckner credits Bickhard with pioneering work on interactions and the processual nature of cognition, I think the connection goes deeper. Bickhard, in his (1993) paper, formulates a theory of content that rings very similar to the ‘forward-looking content,’ in my opinion.

Along similar lines, Lance Nizami expanded this critique into the methodological domain, exposing what can be called the "grand master reader" fallacy in neural decoding studies (Nizami, 2019). Nizami's mighty impressive audit of the literature (he identifies almost 84[4] different postulated mappings of Shannonian posits onto actual neural implementation (op. cit., Appendix Table 1), strongly suggesting a lack of even a hard core within this research program (Lakatos, 1970) or 'research tradition' (Laudan, 1977) reveals that when neuroscientists calculate things like purported "information transmission rate," they are quantifying a relationship that is observer-relative. The neuron itself does not know the set of possible stimuli (the source entropy) and therefore cannot be said to be "reducing uncertainty" in the Shannonian sense. Nizami concludes that the resulting values of information transmission are "artifacts of the observer's statistical power," measuring the experimenter's ability to discriminate stimuli based on neural data rather than the brain's intrinsic processing. By failing to identify a consistent "receiver" within the brain, neuroscience relies on a false fundamental conflation between the substrate and the measure. The tacit assumption on an overseeing external interpreter[5], results in equating the neuron and the scientist's perspectives, leading to the injection of epiphenomenal (or spurious, or, in the strongest terms, fake) "meaning" into physical events.

---

[4] The table is titled *Neuroscientists' Interpretations of Elements of Shannon's General Communication System* and consists of 84 papers starting from Shannon and Weaver's 1949 classic formulation through MacKay & McCulloch (1952) until 2018. Mappings of 5 classical components of Shannon's diagram of information transfer are analyzed: source, transmitter, channel, noise source and receiver. It is a true *tour de force* of philosophical meta-analysis and a powerful weapon to castigate proponents of informational theories into clarifying their accounts. That being said, Nizami's analysis, while being a welcome addition, proves merely that in terms of empirical details the proof is still in the pudding. The final actualization of biological implementations are not necessarily damaging for the IPP as an explanatory strategy in general. I circle back to this idea in the Final Conclusions section, after suggesting some tentative solutions throughout the thesis.

[5] This parallels Godfrey-Smith's critique of the homuncular fallacy in theories of biological representation. He warns that treating neural structures as "maps" creates a dilemma: either the map is just a physical structure, or it implies an intelligent consumer. As he puts it: "If there is a cortical 'map' in the brain, the question of what it is a map of becomes pressing... [This] encourages the idea of a 'reader' of the map... But we must avoid this way of thinking; there is no reader of the map in that sense (Godfrey-Smith 2006:44-5).

The epistemological root of this crisis is traced by Farid Zahnoun to a "semantic trap" (Zahnoun, 2025). Zahnoun argues that contemporary neuroscience has inherited a pre-Shannonian tradition of viewing nerves as wires carrying "messages," onto which it superimposed Shannon's rigorous mathematical theory of communication. The result is the creation of what Zahnoun terms a "mongrel concept" of information—a confused hybrid that conflates the mathematical reduction of uncertainty (entropy) with the transmission of semantic content (meaning). This conceptual drift allows researchers to treat information not as a useful descriptive statistic, but as a literal substance that flows through neural circuits, independent of the physical mechanisms that purportedly carry it.

Despite these withering critiques, IPP remains steadfast. Brette's paper, even though he mobilizes arguments arguably present in the literature at least 20 years prior (Bickhard & Terveen, 1995), instigated a bit of a maelstrom in the intersection between philosophy, psychology, and neuroscience (precisely the humble pasture we are targeting with this dissertation). Plausibly, Brette's empirical credentials (he is a renowned and accomplished practicing neuroscientist in the first place) contributed to the scale of stirring the pot. Irrespective of the sociology of science side of this story, it spurred a series of comments on his paper, jointly published in *Behavioural and Brain Sciences* (2019).

There, one observes a genuine "counterattack" from IPP proponents that strongly suggests abandoning the coding framework may be scientifically premature. For instance, Gallistel argues there that rejecting the code concept leaves cognitive science without a mechanism to explain the temporal continuity of memory (Gallistel, 2019). He posits that the brain must possess an "addressable read/write memory mechanism" to carry information forward in time. Mere synaptic plasticity or dynamic coupling is insufficient to explain how an organism can store a variable (e.g., the location of a food source) and retrieve it hours later to guide action. For

him, the "code" is not a metaphor but an architectural necessity: without a mechanism to explicitly encode, store (write), and retrieve (read) variables, cognitive science would be left without an explanation for how organisms utilize past information to guide future behavior.

Furthermore, other commentaries on Brette's work suggest that the "homunculus" objection may be overstated. Barack and Jaegle (2019) argue that analysis-by-decomposition—including coding analyses—remains conceptually coherent when properly understood, noting that "function can be attributed to the parts of brains" and that decomposition is "indispensable to explanations of brain function that integrate with psychology." They contend that Brette commits a category mistake by conflating the conceptual structure of coding (which involves linear relationships) with the causal structure of implementation (which can involve circular, coupled dynamics). Similarly, Cao and Rathkopf (2019) distinguish between "modest" neural codes that may be learned through plasticity and action policies, and "immodest" codes improperly projected by experimenters. They argue that neurons can develop something "plausibly meaning-like as a consequence of the functional role they gradually learn to play in the overall economy of the brain," suggesting that coding schemes can emerge from dynamic feedback with the environment rather than requiring pre-specified interpretive frameworks. Additionally, Birch and Smortchkova (2019) propose that rather than abandoning coding altogether, neuroscience should employ "an explicit and demanding set of conditions under which a pattern of neural activity genuinely represents an environmental property," including requirements that correlations be exploited in further processing, guide successful action, and allow for misrepresentation. These responses suggest that while loose talk of "coding" may indeed be problematic, more carefully articulated coding frameworks that account for learning, context,

and functional role may yet provide valuable tools for understanding neural representation.

So, clearly, IPP is still an enticing explanatory strategy. But *why?*

2.2 Information Draw

Information is an enticing concept for two main reasons, I think. Firstly, it implies the backing of a well-defined and mathematically rigorous framework of Shannon's information theory (Shannon, 1948, Cover & Thomas, 2006). Secondly, it is an inclusive and purportedly non-assumptive concept. What I mean by that is that information talk appears to be applicable to virtually any given cognitive agent (actual or potential, including the ones whose 'cognitiveness' we would like to in fact assess). Computers surely trade in information, and so do neural nets and robotic agents. Computer engineering is commonly referred to as 'information technology,' and the current crown jewel conference in artificial intelligence is called Neural Information Processing Symposium (NeurIPS, formerly NIPS), after all. At the same time, information talk is pervasive in animal studies, neuroscience, and psychology, which suggests (as I will argue wrongly, or at least simplistically) that there exists a common currency in cognition, a sort of subject matter for cognitive processing that is at some rudimentary level of analysis common with no provisos for all subfields of the interdisciplinary project of cognitive science. The part of the entire conglomerate most prone to this illusion, and hence of most interest to me, is neuroscience.

The information theory (IT) toolkit is indeed used extensively in neuroscientific practice. For instance, Timme and Lapish (2018) suggest the IT mathematical apparatus is a proper means for understanding the collected neural, multivariate data in order to ascertain how the brain encodes, computes, and manipulates (integrates, filters) and sends over information (i.e., identify the information flow)

about the environment. Measures such as mutual information, Shannonian entropy, Kullback-Leibler divergence, or Jensen distance are deemed to be identifiable features of the cognitive substrate. The problems of applications are admittedly there, according to neuroscientists, but they are chiefly of a technical, not fundamentally theoretical or conceptual nature. So certainly, there is a host of challenges for this project, such as effective data binning (how to effectively quantize real-world continuous brain data into a more processable format), the problem of spotty data (we can only record from a tiny fraction of neurons, with processing limits of high-dimensional data being another technical constraint), or finding the best techniques for significance testing. But the questions of how information is fundamentally connected to the emergence of content and effective representation are usually brushed aside. Brains compute information, and information theory is the way to go in understanding this phenomenon.

It has to be admitted that, having accepted these two starting assumptions, what emerges is a fairly enticing picture. We seem to be able to identify the actual substrate of cognitive processes (information) and possess a ready-to-use formal framework to measure and identify it reliably (information theory applied to neural processing). Two additional research programs bolster this view. The first is certainly the artificial neural networks program (Goodfellow, Bengio, Courville, 2016, LeCun, Bengio, Hinton, 2015). Having achieved earth-shattering progress in a growing range of tasks, from visual categorization and autonomous driving to machine translation, celebrated Go mastery and Atari arcade proficiency, the artificial neural networks strongly suggest that the neurally inspired organization of information processing is steadily zeroing in on cracking cognition. This is solidified by the ever-present conception of the very ‘neuro-inspiration’ or ‘biological plausibility’ commonplace in the field (Love, 2021), which suggests that success is connected to the very convergence of ye olde human wetware of

biological information processing and new and shiny artificial processing instantiated in silico.

This point is explicitly put forward in the literature (e.g., Yamins & DiCarlo, 2016, Botvinick et al. 2019), especially in the case of the visual cortex and computer vision algorithms. It is argued that computational models utilize the same organizational principles in visual processing and hence can be fruitfully interpreted as models of human perception. An example of such a strategy is sparse coding (Olshausen & Field, 1991), which originated as a computational principle for effective coding of visual data but has also been found to be implemented in the human visual tract (as deep as the V1 layer). On this basis, Cichy & Kaiser (2019) suggest that ANNs should be treated also as models in a strict, philosophy of science sense as explanatory and predictive tools in understanding vision.

A second scientific success story that lends credibility (at least intuitively) to the idea of an informational crackdown on human brains is brain-computer interfaces (BCIs) and neural decoding programs. BCIs are instrumental setups that are designed for controlling computers not through traditional controllers such as a mouse and keyboard but by brain electrical currents. Usually, this includes specialized hardware for registering the electrical intra-skull signal, software for analyzing it, and translating it into command input for the computer. As such, it is especially promising for people with serious cognitive impairments, such as locked-in patients and people with severe muscular atrophies or neurological damage. Sometimes it also helps to circumvent a faulty neurological circuit, creating a sort of electrical bypass, as in patients with spinal cord damage that renders them unable to control their limbs (Rao et al. 2014). But in principle, it can be used by any individual, also outside the purely medical and therapeutic context. The effect is understandably quite spectacular, delivering prima facie a long-dreamt device controlled purely by the ‘strength of one’s mind,’ straight from

futuristic visions. Think of the 1983 film *Brainhat* starring Christopher Walken or classic William Gibson's world of *Neuromancer*, where the link between neural and artificial information systems is as commonplace as establishing protocols for communication between two computers. If the brain can freely communicate with artificial systems and roam the infospace, and we can establish translational work between them, then the brain (or even the mind) is supposedly prone to being pried open or even 'jacked into' (as Gibsonian lore holds).

One would reply that, well, this is merely literary fantasy, which has little to do with our scientific assessment of what these impressive feats of engineering mean for the theory of cognition, and particularly the information processing thesis. This is, of course, true, but the very vision is also openly embraced in scientific communication. For example, the venerable *Lancet* magazine published its 2011 report on pioneering BCI work under the very title 'Controlling machines with just the power of thought' (Luna, 2011). Respected, general-purpose media outlets keenly follow the surge of excitement with headlines routinely declaring that this kind of technology is 'no longer sci-fi' (as assured by also quite prestigious, if more popular, *Wired* magazine (Rogers, 2020, Pardes, 2020); or BBC (Woollacott, 2018); or *The Economist* (Palmer, 2018)... the list really goes on). What is tacitly assumed is that if technology was proved to be working, the underlying assumption has been vindicated, too. But the technology is concentrated on successful control—generating motor (or quasi-motor in this case) commands to be executed not in vivo, but in silico. Does it really tell us much about the content of information processing in our heads? However, the question springs to mind, how much is then revealed about the content of this newly connected neural activity.

This becomes more apparent as the BCIs shift from fighting cognitive impairment into more paradigmatically 'intellectual' tasks, such as speech and text generation. To give an example, the very recent result of BrainGate researchers (Willett et al.

2021) in mental handwriting has been achieved by instructing subjects to imagine handwriting and, on this basis, teaching the artificial neural algorithm to generate corresponding symbols (letters and words) with high accuracy (94.1%) and at an impressive pace of up to 90 characters per minute (comparable to dexterous digital natives and definitely outpacing some humans—at the very least, my parents). What is interesting here, which follows us into what I hinted at a few paragraphs earlier, is that this empirical success is credited to the ability of recurrent neural networks to do neural decoding of the letters intended by the subjects. Here we arrive at the second, much more bold and interesting (in my opinion) promise of neuroinformational science—the ability to find out the neural code and then, surely, crack it. Following the groundbreaking Haxby et al. (2001) study demonstrating that categorical discriminations can be inferred from neuronal activity, researchers have often demonstrated the ability to purportedly predict at least some rudimentary content of perceptual processing. I say rudimentary as the granularity of the approach has been long limited to categorical discrimination, but recently the much more robust experimental demonstration of this paradigm can be seen in Rashkov et al.'s study (2019), in which the researchers showed that they were able to reconstruct the natural image (i.e., the view in the visual field of the subject) by analyzing the EEG signal data (mostly R300 band) recorded in real time. Similar effects were achieved by Horikawa & Kamitani (2017) with fMRI data in regard to both observed and imagined data (but there the prediction envelops only the classification into categories).

There is also yet another important motivation for sticking with the informational level of analysis, I think. It has to do with the fair methodological stance towards different types of agents and, more precisely, overintellectualizing the problem, hence promoting an anthropocentric perspective on what constitutes certain cognitive powers. It has to be conceded that it goes against important philosophical

traditions of Kantian lineage. We can recall here an important Sellarsian idea of 'two-ply account' (Brandom, 2014) from his hugely influential *Empiricism and Philosophy of Mind,* where he differentiates 'sapients' (true knowers, humans) from 'sentients' (all the merry rest). This non-human bunch is classified as being serviced by 'reliable differential responsive dispositions.' A similar path is taken by, for instance, Carruthers (2006) in respect to animal cognition. We do not want to dwell on this vast subject here, as it does not concern us in this dissertation. But I think it is important to underline the type of philosophical stance that I want to argue against and why we should abandon it and replace it with the empirically minded project of naturalization of information, content, and representation (Neander, 2007, Millikan, 1984, Dretske, 1981).

Firstly, I think it is grossly misleading (and false) to posit that the main (or even only) interesting difference in cognitive bestiary is the human/non-human great divide. It makes little sense to say that primates are not fundamentally dissimilar to thermostats but worlds apart from humans. We do know that animal models are highly accurate (and predictive) in understanding human cortical structures, for example. Macaque cortex has been successfully used to map certain parts of human cortical areas (Tootell et al. 1988; Matsuoko & Komatsu, 2005), and even rodents are routinely used to adequately account for mapping the neural circuits (and assess therapeutic interventions) with direct consequences for studies on humans (such as the groundbreaking work on cognitive maps for navigation in rats; Tolman, 1949; O'Keefe & Dostrovsky, 1971). We also do not want to get rid of evolutionary, phylogenetic explanations that are rendered inadmissible when we do want to erect an explanatory barrier between the species or clades. Such work has already been demonstrated to be fruitful, for example, in spatial navigation tasks (e.g., Morris et al. 1982). But secondly, regardless of actual evidence (which has been mounting ever since the publication of Sellars' 1956 book *Empiricism and the Philosophy of Mind*),

we should avoid this kind of methodological stance for purely philosophical reasons, too. After all, we do not want to declare a particular feature of cognition to be established a priori and then argue for it ex post.

Having said that, we should be wary that it is not by any means equivalent to the fact that the difference will not turn out to be significant (between humans and non-humans, for example). It is simply to say that we want to find them, not declare them. And precisely for this reason, an information-based approach is important as a starting point in the philosophy of cognitive science, as it provides a level playing field for different types of agents. From this foundation, we hope to scaffold more sophisticated explanatory structures, and examine which agents can achieve a certain proficiency (and by virtue of what kind of informational structures they utilize) that the simpler ones lack. Only this kind of approach demystifies cognition and also gives us solid ground to assess artificial agents. This was the approach that guided Dretske in his seminal account of explaining more sophisticated cognitive structures, such as language or knowledge, in informational terms.

Bearing in mind these preliminary and sketchy reasons for answering the general question of why we use *information* in explaining cognitive systems, we can move to the 'what?' and 'how?' conundrums. The first one will lead us into philosophical inquiry into the ontic status of information in cognitive science's practice (and its many shades). The second one will require analyzing how informational structures can bear relation to the next level of the cognitive ladder—representation. In the recent paper Mark Sprevak (2020:591) puts it nicely: 'dating back at least to Dretske (1981), an influential answer [to the question of the relation between information and representation] has been that information is a rung on a ladder that gets one to representation. Representation is information, or representation is information plus some other ingredient'. We need to understand this additional ingredient, perhaps analogously as epistemologists look for the "X" factor in the $JTB + X$

theories of knowledge. To furnish the mind full of information with true representation needs more conditions to be satisfied. After all, the crux of the task can be formulated in Nico Orlandi's catchphrase from their 2014 book (Orlandi, 2014, ch. 3): "while information is ubiquitous, representation is not."

3. Desiderata for Explaining with Information Theories

So, what are we expecting to achieve anyway? To simplify this quite daunting task, it seems reasonable to draw out the desiderata that the acceptable, amended theory (or theories) of information should adhere to. The list is both motivated by the original merits of classical (Shannonian) theory, which were responsible for its popularity and theoretical prowess, but also its shortcomings, especially in the cognitive context. The perfect scenario would be to retain the mathematical rigor, measurability, and great flexibility of Shannonian information but provide it with better connection with problems and notions endemic or crucial for cognitive science, such as content or representation. I doubt that the all-ranging theory that can cover all the desiderata is possible at this point. Still, I think it is sensible to provide a checklist of coveted aspects[6]. Different notions and theories of information can then aspire to fulfill at least some of them with clear demonstration of their own shortcomings. It would provide us with both a good overview of the theoretical landscape, a taxonomy of sorts, but also a good tool for comparing and contrasting rivaling theories. Also, it would hopefully shield us from conflation and misunderstandings, which I argue plague the cognitive sciences

[6] This taxonomic approach aligns with the framework proposed by Winning and Bechtel (2016), who argue that future philosophical progress in this domain requires acknowledging the multi-faceted nature of information. They propose differentiating informational accounts along at least five dimensions: the type of intensional content, the type of vehicle, the content tokening scheme (e.g., causal vs. teleological), potential truth values, and the type of extensional content.

today. Pluralism is then not directly advocated for, but seems to me like a natural conclusion because of how vast the territory that information theories are supposed to cover has become.

It is crucial to qualify the nature of the desiderata that follow. We are not engaged in a metaphysical search for the "essence" of information in the universe (a task for fundamental physics). Rather, our inquiry is pragmatic and naturalistic. We seek a concept of information that is explanatory within the context of cognitive science.

Cognitive science explains behavior by positing internal states that track the environment and guide action. Therefore, any useful definition of information must be "friendly" to this project. A definition might be mathematically rigorous (like Shannon's entropy) yet explanatorily inert if it cannot account for how a specific neural vehicle carries specific semantic content for a downstream consumer. We are looking for the *ontic* properties: real, physical features of the signals, but only insofar as they justify the *explanatory* practice of attributing content to the brain.

Desideratum 1 (henceforth D with numerals): Quantifiability. Theory of information should provide a measure of how to quantitatively measure the amount of given information.

D2: Universality. A good theory of information should be agent-agnostic, that is, should not differentiate 'beforehand' which agents are capable of information processing and which are not. If the given system can operate within a given set of procedures defined by the theory as minimally informational, then it should be considered fair game for comparisons with other capable agents. This requirement is intended to retain the great virtue of original information theory that did not discriminate between animate and inanimate objects or different size and temporal scales as long as they adhered to the coding theorem and other rudimentary (and simple) requirements set by Shannon. For what we want from an

information-based theory is a tool that is less restrictive than traditional philosophical, high-level accounts of cognition that excluded, e.g., non-human agents a priori just by virtue of not possessing the propositional language of practical reason (declared from the armchair). Furthermore, we expect information theory to be a kind of lingua franca for varied sorts of cognitive systems analysis, so we intentionally prefer the theory to be promiscuous in this respect.

D3: Semanticity. The theory should provide a clear statement of whether there can be semantic properties (such as aboutness) extracted purely from the relations defined by it (i.e., informational relations). Preferably some sort of semantics should be present within an information theory. Especially whether informational states can be contentful should be resolved within a theory (as demonstrated by Skyrms, 2010 or Shea, Godfrey-Smith and Cao, 2018). When the semantic relations are originally latent, the theory should be amended to make them explicit. The example of such a case can be seen in Isaac's (2019) extension of the Shannonian paradigm.

D4: Representationality. Somewhat analogously, the theory should clearly testify about its relations with representations. It ought to be made clear whether the information theory itself is supposed to be the theory of representation (as argued by Martinez, 2020, for example) or some other relation is postulated. If the latter is the case, then this relation should be clearly defined—what kind of additional conditions do informational states or informational flow have to meet to ascend to the representational standard?

D5: Leveling. Information theory, at least in the cognitive applications, should take a defined position on Marr's three-level hierarchy. If the theory is supposed to be a computational-level description (as seems most plausible and is usually defended as such), then the boundaries with the two remaining levels need to be delineated. Similarly, the relations between the levels should be specified.

D6: Factivity. The theory in question would be better off if it picked a side in the factivity debate. The possibility of false information should be openly allowed, or excluded. If the latter is preferred, an explanation for phenomena of misinformation and disinformation (as separate kinds of faulty information) should get a different kind of conceptualization.

D7: Objectivity or receiver relativity. Satisfactory treatment of information ought to have a clear answer as to whether it is aimed at providing a description of things 'out there,' independent fabric (the commodity view of information that is extractable from the environment) of the external world, to which different receivers have equal access or not. If not, the theory should specify how the receiver-relativity phenomenon can be addressed without jeopardizing the universality criteria (as stated in D1 and 2).

D8: Internality/externality. The theory should make clear commitments to whether it can differentiate between the information within (for) the system and information about the system (i.e., to the observer or experimenter). In another sense, it should tackle the problem of consumers of information, whether the informational content. For example, is the next neuron in the brain-signal relay in the same position as the experimenter spying on it when both perform a 'readout' from a previous neuron?

D9: Arbitrariness. The theory should specify if the informational relations have the nature of a purely arbitrary, conventional code (or at least in principle could have, even if not realized in implementation) or not. Or conversely, if some kind of natural, pragmatic, or otherwise preferable mapping (coding) is assumed, it should be clearly stated, i.e., the motivations of such bias and its structure should be offered.

D10: Epistemic closure. The theory should specify whether there are logical relations between pieces of information that hold relative to the given subject's possession of such pieces of information.

D11: Uncertainty. Information theory should explain how it relates to the question of uncertainty. Is informing in the business of uncertainty reduction, as in the original formulations, or is there some other relation assumed[7].

Such a list is probably not exhaustive, nor is it to be taken en masse (as I mentioned before introducing it). Nevertheless, it is offered in this thesis as a model against which we should be testing and assessing rivaling accounts of information, especially in the cognitive context of the term.

4. From Information Theory to Theories of Information

Now I would like to turn to reviewing the most influential theories of information to highlight their differences. As Piccinini and Scarantino observed, "when the cybernetic movement became influential in psychology, AI, and neuroscience, 'computation' and 'information' became ubiquitous buzzwords" (2010:244). Neural contexts are especially prone to the usage of information-talk in quite a willy-nilly way, as a sort of intuitively obvious 'semantic analogue of matter,' which is simultaneously hidden in the environment to be mined from and then flowing (somewhat liquid-like) within the brain. Such a practice might not necessarily be detrimental if it were restricted to informal contexts. All trades use jargon to communicate more freely and flexibly, after all, and neuroscience is a trade like any

[7] Going back to Winning and Bechtel's (op. cit.) proposal, my list of the desiderata (D1–D11) can be viewed as an expansion of their requirements tailored to cognitive neuroscience; their dimension of 'intensional content' maps to my D3 (Semanticity); their 'type of vehicle' roughly corresponds to D4 (Representationality); their 'content tokening scheme' underwrites D7 (Objectivity) and D9 (Arbitrariness); and their 'truth values' dimension is directly captured by D6 (Factivity).

other. Unfortunately, the casual usage permeates scientific communication, which significantly obfuscates the nontrivial nature of neural information processing (Baggio, 2017). But it is conceptually dubious to imply semantic function while at the same time one holds that it is an easily operationalizable (measurable and spatio-temporally tractable) quantity, as neuroscientists routinely do (e.g., Blinowska et al.,2004; Borst & Theunissen, 1999).

This tension reflects a feeling of broader conceptual uncertainty that is palpable across the literature. To paraphrase Griffiths (2001) who famously noted in the context of biology, information is still very much 'a metaphor in search of a theory.

4.1 Classical (Shannonian) Theory of Information

Shannon's theory (Shannon, 1948; Shannon & Weaver, 1949) is a paradigmatic formal, mathematical treatment of information. It provides a clear definition of information transfer in terms of sources, channels, receivers, and noise sources. It operates on rigid operations of encoding and decoding of messages and aims at a quantitative description—measuring the level of entropy understood as a reduction of uncertainty through the transmission of messages. Furthermore, it requires the set of permissible symbols to be selected from with some given probabilities of occurrence of each symbol. Shannon's theory has two important features: it is intentionally very abstract and aimed at solving a particular engineering problem, namely a problem of optimal communication. This optimal communication means the best way to encode, send, and decode a given signal, or, in other words, transmit the signal from sender to receiver. It defines the maximal efficacy (upper bound) of the transmission (the capacity of the channel) and demonstrates the condition for optimal communication when considering the nature of the underlying set of messages, the level of noise, the threshold of redundancy, etc.

According to the founding father of the theory himself, the aim is to solve in this way 'a fundamental problem of communication.'

This problem is simply to send a message produced (selected from the pool of possible sets of signs, i.e., messages) at a source and recovered at a destination (picked the same/similar-enough concatenation of signs). Then it amounts to mapping the original message to the received one so as to satisfy relevant desiderata (such as speed, cost, or accuracy). 'The sense in which the semantic aspects are irrelevant is precisely that messages carry information merely qua selectable, physically distinguishable structures associated with given probabilities' (Piccinini & Scarantino, 2010: 240). Because it is a purely abstract framework, it can be applied to a virtually endless class of problems, as long as the signal can be described probabilistically in a regimented manner. Furthermore, the connection between the meaning of the message and the code is purely arbitrary and not necessarily connected by any physical relation other than rigid encoding. A string of binary digits (the most popular, but by any means not the only possible format for Shannonian coding) does not bear any relation to its 'target' meaning (i.e., a word or an image). The word 'Shannon' or the image of Charles Shannon himself bears no similarity to the zeroes and ones that encode it in the memory of the computer I am writing this chapter on. But after all, so do linguistic structures (as we said, the 'shape' or sound of the word does not resemble its designation). The difference is, however, the fact that at least on grounds of traditional interpretation, informational states of Shannon's do not convey meaning in the traditional sense at all.

That's why informational states are supposed to be not about things as such, unable to contribute to the meaning. The orthodox interpretation of information theory seems in line with this assertion. As famously put by Freeman Dyson to information theory, 'meaning is irrelevant' (Dyson, 2011; per Shea et al.,2017:1009). Shannon

himself was openly skeptical about the extensions of his theory to anything more than a mathematical theory of effective communication and formal description of signaling channels. He considered the ascription of meaning to information theory to be a fundamental misuse of its core function and attacked it in an article forthrightly titled ‘The Bandwagon’ (Shannon, 1956).

But that did not stop the other theorists from continuing to run with Shannon’s definitions and equations all the way to the bank in their respective disciplines. In neuroscience, the publication of McCulloch and Pitts’ (1943) study on the logical calculus of nervous signals cemented the idea that neurons are digital, binary transmitters of information (can be ON or OFF, depending on whether the spike occurs or not). McCulloch’s cooperation with von Neumann (Pierce, 1980) and his ideas of formally defining the computing machines as information processors was one of the main fuels of the conviction that brains are fundamentally computing machines trading in and out the same kind of Shannonian information (Piccinini, 2004). Shannon information is still routinely used by neuroscientists to measure the quantity of information carried by neural signals about a stimulus and then measure how efficient the respective coding is (in other words, which neural patterns fit best to optimal information carriage about the stimulus; Dayan & Abbott, 2001).

While critics often dismiss this application as a “vacuous metaphor” (e.g., Epstein, 2016), proponents like Kevin Lande (2019) argue that the statement “the brain is a computer” is a literal, scientific hypothesis that generates precise predictions. Lande contends that the IPP offers a formula: "The brain computes X." By filling in the variable X with specific mathematical functions, scientists generate testable hypotheses validated by neurophysiological evidence. Unlike the "empty metaphor" criticized by detractors, Lande illustrates that cognitive science has successfully identified specific, empirically detectable computations:

- Direction from interaural time difference: When you locate a sound, your auditory system is literally computing an estimate of direction based on the difference in arrival times of sound waves at your left and right ears, utilizing trigonometric principles.
- Depth from disparity: The visual system computes depth by calculating the "disparity"—the difference in the relative positions of retinal cells responding to the same point in the world.
- Surface color from illumination: To perceive a stable object color despite changing light, the visual system computes an estimate of the object's reflectance by mathematically discounting the estimated illumination.

These are not metaphors; they are concrete mathematical operations implemented by the physical wiring of the nervous system. As Lande asserts, "Nobody has any real idea of how else to understand our abilities" without these specific computational hypotheses.

4.2 Algorithmic Information Theory (AIT)

Shannon's theory received an extension in the form of Kolmogorov-Solomonoff-Chaitin's interesting algorithmic interpretation, dubbed the Algorithmic Information Theory (AIT). It builds upon the fundamental observation of Shannon that information is the reduction of uncertainty about the source. The bigger the reduction (the difference between the probability of the 'expected' and really received signal), the more information was then conveyed. Intuitively, when we know that in the transmission of letters, 'E' is by far the most common (has the highest underlying probability), then receiving an 'E' is the least informative ('could be expected'). Scrabble is based on a similar idea, ascribing points to less common letters. So, the Shannonian observation was that the optimal

solution when devising the code was to ascribe the more probable symbols to a shorter description (it is the idea behind Huffman encoding trees).

AIT's main idea was to connect the Shannonian train of thought on minimal (on average) description length with the observations on randomness, compression. The simplest recapitulation of the main idea states that compression is a process that encodes the original piece of information (signal) with a shorter string (lower bitrate) than the original formulation. This kind of operation can only succeed when there is an element of redundancy in the signal to be encoded. Redundancy can appear only if some kind of regularity organizes the given signal or dataset. Entirely random processes are then fundamentally uncompressible. This crucial relation served Chaitin (1977) to define the very notion of randomness in terms of the compressibility of its respective signal. Put plainly, if there is no possibility of representing the original message in a shorter form (or, in other words, generating a shorter-than-original program code), then it is entirely random (Dennett, 1991). This line of research was extremely prolific in computer science, and by extension in convincing scientists who wanted to transplant computational principles into cognitive science that effective abstraction or generalization of data, understood as looking for patterns within it, has to do with arriving at more compressed representations of the data in question. We will return to the issue of this intricate formal relation between compressing and representing later on.

4.3 Semantic Theories

However enticing, the formal theories run into apparent trouble when we try to understand the semantic aspect of information processing. Let's recall the observation that, according to Shannon's theory, the most improbable strings generate the most information for the agent. When we interpret this mathematical problem in the light of linguistic symbols, we clearly see that a problem emerges, as

most nonsensical concatenations of symbols would be most informationally rich. From the mathematical point of view, "@#d!!DfCCf9vosc" is an equally admissible string as "armadillo." Yet, because of its randomness and usage of rarer symbols, it yields more surprises. In some cases the mathematical strength even becomes a practical one, as is the case for setting digital passwords. So, the gist of that relation needn't be purely abstract.

This problem can be exacerbated even further. If we accept that a true statement has a probability of 1, then the transmission of it does not shift the probabilities of the receiver, rendering it informationally meaningless. Conversely, impossible statements would gather maximal informativeness, as their surprise value would be the highest. This observation has been made by Bar-Hillel and Carnap (1953), and hence this has been known as the Bar-Hillel-Carnap paradox. It is possible to circumvent the problem by stating that the $P = 1$ ascription is justified only for necessarily true statements. Other contingently true statements would have a lower (maybe still high, but non-1) probability for the receiver prior to the acceptance. But such a move introduces different values of the state relative to the temporal history of the receiver, suggesting an epistemic or 'subjective' interpretation of information, which is at odds with the classical reading of information as based on values that are, in a sense, externally fixed.

A similar point has been put forward by Hintikka (1970) with regard to deductive reasoning. If valid and sound inference guarantees the transfer of logical value, then there is no probability-shifting procedure, no reduction of uncertainty in objective terms. Hence, it seems to follow that any deductive reasoning has an information value of zero. This conclusion is undoubtedly deeply troubling. No wonder that Hintikka decided to dub the problem 'the scandal of deduction,' similarly to Humean (1748/2007) treatment of induction (Sequioah-Grayson, 2007).

These kinds of problems motivated theorists to devise an account of information that would take meaning into account. For we are intuitively interested in a theory of informativeness as a function of what the signals stand for. On this view, the informativeness now depends also on what particular sign stands for, or in other words, which feature or state of affairs it carries information about (contrary to simply how comparably likely other alternatives were statistically).

Semantic accounts usually face an important theoretical problem when they try to achieve that goal. Intuitively, there should be a difference between states of affairs that are connected by physical laws and by convention, which should be acknowledged within the theory. The famous motivation for that was given by Grice (1957), where he delineated conditions for differentiating natural and non-natural (conventional) meaning. In a famous set of examples, he illustrates his point:

“Those spots mean (meant) measles.”

Contrasted to:

“Those three rings on the bell (of the bus) mean that the ‘bus is full’.”

The first example is supposed to indicate a situation in which the information-carrying vehicle (spots) informs about the cause that is linked to them in virtue of natural law. Hence, it does not seem plausible to utter a sentence like “these spots mean measles... but it is not measles.” The sentence S means (in a natural sense) p entails that p is true. If the smoke means fire, it logically follows that there is a fire. In the second case, however, this does not obtain. As the relation that ties the bell sounds with the property of the bus being full is non-nomic, it does not logically follow that what p means by S is necessarily true.

This distinction has been hugely influential in trying to pry apart truly intentional signaling from natural signaling. But this dichotomy has been recently repeatedly

attacked. For instance, in animal cognition research, the practitioners called for ‘exorcising the Grice’s ghost’ (Townsend et al.,2016), arguing that holding dead the Gricean divide, one obfuscates the true differences between human and non-human actors by arbitrary decree (of what should in fact be studied empirically). But the Gricean project has also been undermined on philosophical grounds. Most notably, Brian Skryms, in his groundbreaking 1990 book *Signals: Evolution, Learning and Information* declares right off the bat that ‘all meaning is natural meaning’ and proceeds to present the theory of the emergence of conventional signaling from interplay between senders and receivers in nature. Following David Lewis’ 1969 signaling games from his PhD dissertation *Conventions’* Skyrms offers a program of reducing the non-natural meaning to natural meaning. Similarly, Godfrey-Smith, Shea, and Cao (2017) suggest an alternative formal strategy of explaining the emergence of content through the dynamics in simple signaling systems. More on that to follow in the chapter on informational theories of content.

4.4 Ecological (Gibsonian) Information

Another, very influential and still vibrant theory of what should count as information and what its role is in cognition is found in the tradition of ecological psychology, originating with J.J. Gibson (1979) and recently utilized and refined by theorists such as van Dijk, Withagen, and Bongers (2015), thus sometimes dubbed as neo-Gibsonianism. Gibsonianism, whether in its traditional or updated version, constitutes a distinct and quite radical alternative to the computationalist orthodoxy. While the standard information processing paradigm views information as an internal commodity that must be inferred, processed, or enriched by internal mechanisms to become useful, the ecological view argues that information exists objectively in the environment, available to be "picked up" directly by an attuned organism. This perspective fundamentally challenges the "poverty of the stimulus"

dogma that underpins much of cognitive science, positing instead a "richness of the stimulus" that renders internal representation unnecessary for adaptive behavior.

The core of the ecological theory is the distinction between mere correlation (or covariance) and specification. Standard informational theories—such as Shannon's mathematical theory or Dretske's causal-informational account—rely on probabilistic correlations: a signal r indicates a source s because they co-vary reliably $P(s \mid r) > P(s)$. However, correlations are rarely perfect; they leave room for uncertainty (equivocation), necessitating internal guesswork or inference to reconstruct the cause.

In contrast, ecological information is traditionally defined by lawfulness—a 1:1 relation between the ambient energy array and the environment. However, van Dijk et al. (2015) warn that relying solely on this environmental correspondence (often phrased as "information about") is insufficient to escape the charge of content. If information is defined merely by the array "carrying" a specification of the world independent of the animal, it invites the view that it is a message waiting to be decoded. Instead, they argue that Gibson's mature view was that the array becomes information only through its use by an active animal. By shifting from "information about" to "information for affordances," they posit that information is not constituted by static laws alone, but by the animal's active coordination with the environment.

This usage-based definition leads to the conclusion that ecological information is truly information without content. In the semantic tradition, content requires conditions of satisfaction—the possibility of being true or false. However, van Dijk et al. argue that because information is defined by its pragmatic function, it "cannot be evaluated as being more or less true or accurately corresponding to an affordance – there are no conditions to satisfy it being about the affordances".

Consequently, ecological information implies no truth conditions; "rather information can be more or less useful for adapting to the environment, that is all". It informs the organism not by carrying a message about the world, but by allowing the animal to "adaptively coordinate with its world" (op. cit.; p. 213).

This view is echoed by Turvey (2018), who argues that the "information" in ecological psychology is a physical constraint, akin to a gear in a machine, rather than a symbolic code. The "meaning" of an affordance (e.g., "climbability") is not a property of a mental representation but a property of the organism-environment system itself.

Importantly, this non-representational view is explicitly conjured by empirical researchers in cognitive science, especially in motor-control studies on problems such as dynamic touch or postural control. For instance, Doyon et al. (2019) investigate the perception of "stand-on-ability" (an affordance) and demonstrate that it is modulated by the multifractality of postural sway. The organism does not "calculate" the surface angle or friction; rather, the subtle, chaotic fluctuations of the body allows it to detect invariant structures in the global array that specify the affordance directly.

Similarly, Stafford et al. (2019) analyze safe road-crossing behavior (the "gap acceptance" task). They show that successful crossing does not depend on an internal model of car velocity and distance. Instead, pedestrians "tune into" specificational optical variables that define the "safe gap" directly. The behavior is emergent from the coupling of the agent's movement and the optic flow, without any intermediate step of creating a "snapshot" or representation of the scene.

However, treating the ecological camp as a monolith would be a mistake. As recent analyses reveal, the 'Neo-Gibsonian' landscape is fractured into distinct factions regarding the status of semantic content. On one side stand the 'purists' like van

Dijk et al., who argue that Gibson's mature view shifted strictly to 'information for' affordances—a pragmatic category that admits of utility but not truth or falsity, effectively purging the system of semantic content (by controversially equating all content to truth-evaluable content, which make imperative and instructional content conceptually impossible).

On the other side are what we might call the 'squirmers' (e.g., Chemero, de Oliveira). These theorists often employ definitions of information (based on correlation and 'aboutness') that technically meet the formal criteria for representation. Yet, they evade the label by adopting 'artifactualism'[8] —arguing that while their internal states could be described *as* representations, this status is incidental to their primary function as pragmatic tools for action. This maneuver often rests on the questionable assumption that 'true' representation must be exclusively declarative or propositional, ignoring, for example, Millikan's (1995) insight into pushmi-pullyu representations that are simultaneously descriptive and prescriptive.

This tension reaches a peak in the work of Kiverstein and Rietveld (2018), who attempt to explain higher-order cognition via 'virtual affordances.' While they deny representationalism, their concept of engaging with 'non-actualized possibilities' implicitly smuggles in conditions of satisfaction: to successfully 'pretend' or anticipate a virtual state requires a normative standard of correctness that functions functionally identically to truth conditions.

Finally, it is crucial to correct the common misconception that ecological information is purely qualitative. Grounded in dynamical systems theory, the ecological approach is rigorously quantitative. As Lee (2009) demonstrates via his

---

[8] Intuitively, we can think of it as "fictionalism', but they adopt the term because they are concerned with models and scientific *artifacts* in general (e.g. Sanches de Oliveira, 2002).

tau theory, ecological information can be defined by precise geometric invariants and dynamical metrics (such as time-to-contact ratios). Alicea et al. (2024) recently proposed a Gibsonian Information (GI) measure—a mathematical model that quantifies an agent's spatiotemporal sensory input. Thus, the rejection of the 'computational' label by Gibsonians should not be confused with a rejection of mathematical apparatus and the willingness to quantify information; their variables could in principle be as quantitative as any in the IPP, even if their ontology differs.

Admittedly, for the purposes of our inquiry, the ecological view presents a formidable challenge. It satisfies the objectivity challenge perfectly: information is a real, physical feature of the environment, available to but independent of the observer. It also satisfies the structure desideratum, acknowledging the rich, regimented, but invariant geometry of the signal.

However, Gibsonianism struggles because, by denying the existence of internal vehicles, it confines information to the immediate interaction. As Chemero (2009) notes, the ecological agent is perfectly attuned to the "here and now" but structurally incapable of thinking about the "there and then" (offline cognition). Without an internal vehicle to "carry" the information away from the specificational source, the system is causally tethered to the immediate stimulus (hence the allure of virtual affordances). Furthermore, by defining information as lawful specification, it accepts that information cannot misrepresent. This renders it insufficient for explaining high-level cognitive errors, hallucinations, or the generativity of imagination, which require a vehicle that can stand *in* for the world even when the world is not cooperating.

4.5 Integrated (IIT) and effective information (EIT) theories

Recently, another influential theory of information was developed by Tononi (2004) in cooperation with Sporns (2003) and Koch (Tononi et al. 2016), among others. It

was directly inspired by work on sleep and awareness and, as such, geared mainly at providing rigid measures of consciousness. They were embittered by the fact the study of consciousness stalled, and their opinion got bogged down into hand-wavy, barren debates. They thought the way to break the deadlock was to look for a quantitative measure of consciousness. In an extremely crude sketch, it was supposed to provide a yardstick of increasing awareness. In their view, it is the amount of integrated information that provides such a measure (they decline the identity interpretation, where integrated informational states are supposed to be conscious states in a literal way). The more integrated consciousness, the more consciousness it can be declared to possess. To do so, Tononi and Koch suggested introducing a different measure than Shannon's, a theoretical object called 'phi' (Φ).

Without going into too much detail that does not concern us here, suffice to say that IIT as a theory of consciousness entails 5 core principles: intrinsic existence, composition, information, integration, and composition. If the system possesses all 5, it can be purportedly classified as a conscious one. While IIT has some backing, it remains a wildly controversial theory. It was accused of being unspecified (Pautz, 2018), leading to absurd predictions (Aaronson, 2014), meaningless (in the virtue of being panpsychic; Searle, 2013), and not empirically testable (as claimed in the open letter to the NIH Brain Initiative by more than 50 scholars in cognitive science). There are also clear implementation problems with the main measure of the theory, as it is not computationally tractable.

What is interesting for our purposes, though, is the claim that IIT proponents suggest a new interpretation of information. For instance, Dennett posits that it "uses Shannon information theory in a novel way" (Dennett, 2017:114). So what does it mean for the information to be integrated? Integrated information 'quantifies to what extent the cause-effect structure specified by a system's elements changes if the system is partitioned (cut or reduced) along its minimum partition (the one that

makes the least difference)' (Tononi, 2015). In other words, informally it is a tool conceived as a measure of how much causal power a particular system loses if it is divided in the most unobtrusive manner possible. It becomes clear then that the aim of the theory is the explanation of causal emergence within complex systems. IIT proponents claim that the description at the macro level is not equal to the sum of micro-level descriptions. IT is supposed to provide a formalism for demonstrating this phenomenon and providing a mathematical concept to rigidly define it.

This overarching motivation and main idea was taken further and generalized by the other within the IIT paradigm. What followed was an extension of the idea in the form of effective information as a measure for causal emergence (Hoel, Albantakis & Tononi, 2013; Hoel, 2018). EI focuses on network science, and the measure is computed on graphs. EI is supposed to assess how much information is stored in the relationships between the nodes. Networks with higher effective information contain more information in the relationships between nodes. The formula consists of two terms. The first one is obtained by taking the average uncertainty of a network node's outputs, which is the Shannon entropy of its out-weights, $H(Wi^{out})$. This average $(H(Wi^{out}))$ reflects the amount of noise inside the network and represents how deterministic the given network is. When $(H(Wi^{out}))$ approaches zero, the network becomes more and more deterministic; if it equals one, it is said to be fully deterministic. The second term is how weight is distributed across the whole network, (Wi^{out}). This vector is composed of elements that are the sum of the in-weights w_{ji} to each node v_i from each of its incoming neighbors, v_j (then normalized by the total weight of the network). Its entropy, $H((Wi^{out}))$, reflects how certainty is distributed across the network. That property is technically called degeneracy of the network. Intuitively speaking, the more degenerate network is when more of all paths lead to the same node. When

$H((Wi^{out}))$ equals zero, then the network is said to be fully degenerate. Degeneracy is the amount of information in the connectivity lost via an overlap in input weights (e.g., if multiple nodes output to the same node). So conversely, in a perfectly nondegenerate system where all nodes have equal input weights and the degeneracy measure is zero, and $H((Wi^{out}))$ is equal to the logarithm of base 2 from the number of nodes. EI is computed by subtraction of these two terms.

$$EI = H(\langle W_i^{out} \rangle) - \langle H(W_i^{out}) \rangle$$

Or, informally, it is a subtraction of measures of degeneracy from determinism. Interestingly enough, the alleged outreach of such formalism is unbounded. According to the authors, it can be used to measure networks of arbitrary complexity and scale, from the sub-personal (even molecular) to the multi-agent ones. As such, EI was mobilized in the debate on downward causation—the idea that higher-level properties are causally active on lower level parts and properties (Flack, 2017). According to Flack, adaptive systems can causally influence their subsystems downward through the process called coarse-graining. Coarse graining assumes computing one's own internal systems in such a manner that low-level descriptions are 'smoothed out' (Flack, 2017:3). The classic example is temperature, where the high-level description of the system (the readout of the measurement) is an average speed of molecule movements of the measured matter. Said differently, coarse-graining is "integration over component behavior." But this regularity extraction is supposedly different from other procedures of its kind, such as lossless compression, in which noise and redundancies are removed from the dataset but where the compressed output is not itself the result of an integration (Flack, 2017:4).

We can perhaps see now why EI is a prime candidate for examination in neural context—network science, causal emergence, and roots in consciousness theory all

suggest the natural connection. Indeed, Klein and Hoel (2020) point out that neural data is a prime field for implementation of EI ideas. Hence, I will revisit their promises and claims in the further sections where I assess the potency of given theories of information in explaining neural processing.

4.6 Dretske's Theory

Perhaps the most influential theory of semantic information has been defended by Dretske (1981; 1984; 1986; 1988). The aim of his account was to retain the original spirit of Shannonian theory of information (or rather its enlargements, such as Wiener's project of cybernetics) as the objective commodity that was ready to be 'mined from' the environment. Dretske also saw an opportunity to use this commodity as a building block for scaffolding knowledge, perhaps analogously to the relation between concepts of matter and physical objects. To use his words, 'information is that commodity capable of yielding knowledge, and what information a signal carries is what we can learn from it' (1981:44). Information was supposed to help in the grand project of naturalizing traditional epistemology. To quote Dretske again, from the preface from *Knowledge and Flow of Information*: 'The entire project can be viewed as an exercise in naturalism—or, if you prefer, materialistic metaphysics. Can you bake a mental cake using only physical yeast and flour? The argument is that you can. ' (1981:xi). He was impressed with the notion that promised to be not only objective and mind-independent but also quantifiable. According to his definition:

'Informational content: A signal r carries the information that s is F = The conditional probability of s's being F, given r (and k, stipulated as background knowledge), is 1 (but, given k alone, less than 1)' (1981:65)

Let's unpack it using an example. What does it mean when the yellow petrol icon springs up on the car's dashboard? An icon is the signal ('r') that carries the

information that the fuel level in the tank is low (F), where the probability of the low fuel level is 1, given the yellow petrol icon appeared (provided the background knowledge of the subject – k). This second condition means that the information-carrying relation obtains only if the F's probability is lower than 1, given only the background knowledge. In other words, r is informational about F, only when it shifts the probabilities, relative to the knower's knowledge, to 1.

This seems to be a very high bar to clear to count as information. After all, very rarely (and supposedly even almost never) we seem to have grounds to suspect the probability of some world-state's occurrence is precisely 1. Another potential problem for Dretske's definition is the consequence of some form of logical closure that follows from his definition. If r carries, relative to the subject S, the information that p, and S knows that p entails q, then r carries, relative to S and k, the information that q (Jager, 2004:192). This is problematic because it suggests that particular signals are more informational than we would like to accept (in other words, informational states are prima facie not closed under entailment). A signal about some state would then logically inform not only on the state (x) itself but also, for example, about the states that are excluded by that state of affairs (D'Alfonso, 2003).

Interestingly enough, this worry is somewhat addressed by Dretske. His solution is to introduce the digital-analog distinction, but in a slightly unorthodox (in his own words) way. I quote him in full here:

'I will say that a signal (structure, event, state) carries the information that s is F in digital form if and only if the signal carries no additional information about s, no information that is not already nested in s's being F. If the signal does carry additional information about s, information that is not nested in s's being F, then I shall say that the signal carries this information in analog form. When a signal

carries the information that s is F in analog form, the signal always carries more specific, more determinate, information about s than that it is F. Every signal carries information in both analog and digital form. The most specific piece of information the signal carries (about s) is the only piece of information it carries (about s) in digital form. All other information (about s) is coded in analog form. ‘ (Dretske, 1981:26)

Dretske then invites us to contrast statements with pictures. According to him, the statements of the form ‘there is coffee in the picture’ ultimately provide only one ‘piece’ of information—namely, that there is coffee depicted. Nothing more is disclosed, no additional information about its color, whether it is in a cup, a mug, or a jar, how it is positioned, and so forth. Conversely, pictures are analog representations insofar as they provide richer informational input (‘nested’ in the original representation, using his terminology). We can reliably infer more properties (i.e., extract more information) about the depicted coffee, which were inaccessible to us on the basis of digital representation. Analog representation is inherently richer and more specific, on his account. On the other hand, digital form is most determinate, distinct. What follows is a fascinating discussion of a hypothetical machine (the ‘whole-template matching device’ recognizing whether the presented input exactly matches the stored pattern), something we would now call a classifier (albeit an extremely primitive one). Dretske thinks that the process of transitioning from complex, rich pictorial representation (the image is such and such) to the simplest digital representation (there is a match) is a process of loss of information. This somewhat overlooked idea of Dretske’s is worth noting here, as I will return to it in the later phases of the dissertation.

After this prolonged digression, I want to return to the discussion of the core of Dretske’s ideas on information. We can recall that Dretske’s definition takes advantage of relativization to the subject S (the ‘knower’) and its background

knowledge. This introduction of the knower into the definition subjected Dretske to significant criticism. It seems that by doing that, Dretske is simply begging the question in trying to explain the emergence of intentional signaling by referring to the notion of knowledge that is already intentional in the first place. This is superficially a valid criticism, but I think that Dretske can be interpreted here a bit more charitably. According to him, the 'knowledge' refers to the state of the receiver, insofar as it is aware of the possibilities at the source. In a sense, this scope of knowledge is limited only to the constraints on the range of possibilities that are subject to transmission and does not account for a more full-blown consignment to propositional-like knowledge on the part of the receiver.

An important and often misunderstood (or misinterpreted) part of Dretskean theory is its purported basis in causal processes in ascertaining whether information flow occurs or not. But this is a mistake. Dretske explicitly denies that this is the case and discusses cases when causal diagrams do not align with the informational flow. This is because the amount of information generated at the source is not necessarily identical to the amount of information carried by the signal. This situation sometimes obtains, but only in special cases where the given value of the sender state ('s') gives rise to a particular receiver state ('r'). In other words, there is no possibility for equivocation of the signal.

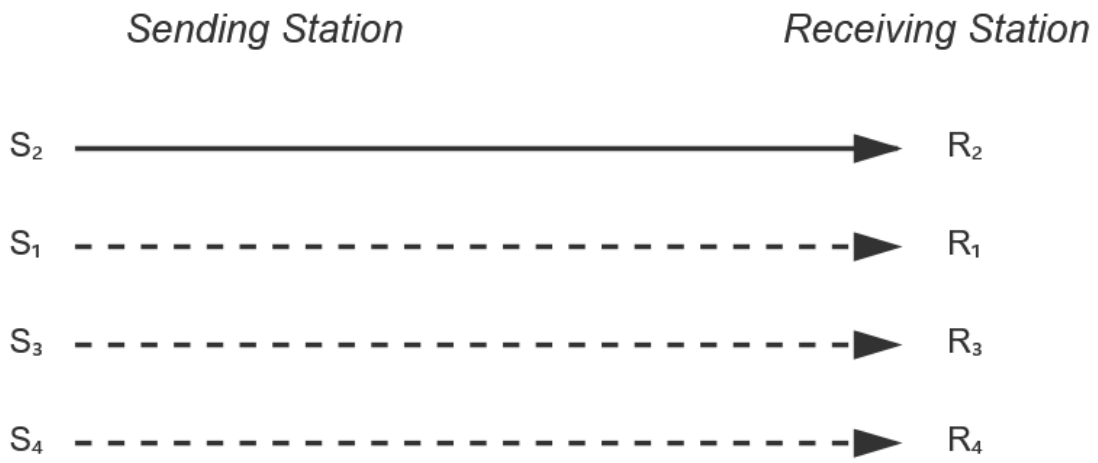

Figure 1: Unequivocal signal transmission (adapted from Dretske, 1981:28)

The dashed arrows represent counterfactual scenarios, whereas the solid arrow represents the actually obtained scenario. Hence, in the case of Figure 1, this problem is not visible yet, as there is no possibility for equivocation in receiver states (r1-r4). But let's take a look at another diagram.

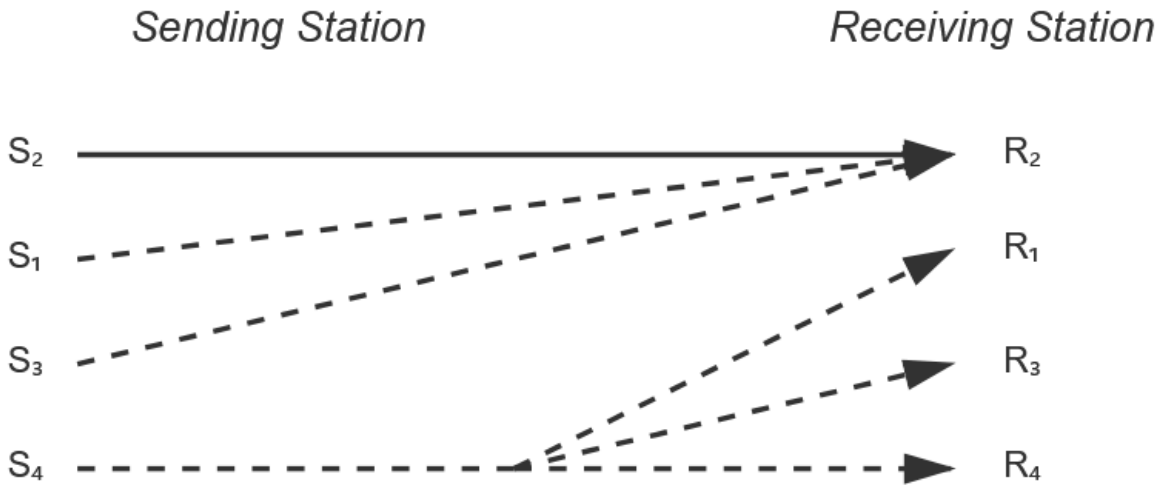

Figure 2: Equivocal signal transmission (ibidem)

Here the causal ramifications are more complex. It is crucial that different states of the source (which are here considered to be equiprobable) could in principle result in the same state of the receiver (here – r2). Then, the informational value of the state on the receiving end is quantitatively different from that informational value of the corresponding sender state (all carry 2 bits, as they split the probabilities equally). But the informational value of r2 is lower[9]. We should bear in mind that from a causal, observational standpoint, this situation would be indistinguishable from the previous one, as only one, same type of causal link obtains 'in the real world.' But as we see, from an informational standpoint, the situation is quite different.

That is why the causal picture itself is not sufficient to ascertain the informational relations between the source and the receiver. We need to consider all possible combinations of counterfactual scenarios to adequately measure the information flow. Dretske concludes this as follows: 'The causal story does not tell us how much information is transmitted because it fails to tell us whether the causal sequence is embedded in a network of possibilities similar to that depicted in Figure 1 or Figure 2'[10] (1981:29). There is yet another problem with tying an informational picture with a causal one. It has been taken to the fore by Barwise and Seligman in the preliminary discussion of their logical framework of information flow. They observe that information disobeys the arrow of time. We can gather reliable information about past states routinely and effortlessly. Otherwise, historians and archaeologists would have a hard time justifying their craft. Any credible information I might have on Harald Bluetooth or the Battle of Hastings is well-formed information instantiated in the present about past states of affairs. The

---

[9] The actual informational value (i.e., how much information it carries) of the receiver state is defined as an informational value of the responsible sender state minus the equivocation (of the given signal).

[10] In numeration adjusted to one used here. In Dretske's text they are labeled 1.6 and 1.7, respectively.

causal picture is, of course, highly restricted in this sense. According to almost all theories of causation, any kind of retroactive causal chain is strictly profane. This feature of 'arrow of time disobedience' was used by some theorists to redefine information, most notably by a recent account of Colgate and Ziock (2011)[11]. Barwise and Seligman did not think of that as a problem but as a feature to be accounted for, so they aimed at devising a formalism to be able to escape the causal connection. They did not, however, follow Dretske too closely.

4.7 Miłkowski's Correspondence Theory of Semantic Information

Dretske's proposal is still, however, covariational. This constituting feature of the theory is not unproblematic, though. Problems with covariance-based theory were criticized right after the publishing of *Knowledge and Information Flow*. The most influential critique came from the teleosemantic camp and continued on for a number of years (Millikan, 1984; 2004; 2017; Jacob, 2000, Neander, 2004, Shea, 2007). The gravity of the problem his book faced was recognized by Dretske himself, who then augmented his original account based on mere indicative properties of informing with teleosemantic ideas, chiefly the problems of information consumption. But recently, instead of following the teleosemantic path, a novel approach to defining semantic information in terms of structural correspondence has been proposed by Miłkowski (2023). Building on Barwise and Seligman's ideas on infomorphism from their monumental book *Information Flow: The Logic of Distributed Systems* (1997) structural correspondence is conceptualized by the relation of similarity that gives rise to satisfaction conditions realized by the relation of infocorrespondence between token-type classifications. This proposal is

[11] While some theorists, notably Bergstrom and Rosvall (2011), have attempted to redefine biological information strictly via transmission to preserve this unidirectionality (arguing that genetic information flows only from genotype to phenotype), such restrictions are ill-suited for cognitive science. Unlike genetic transmission, cognitive information processing must explain the retrieval of past states (memory) and the anticipation of future ones (planning), necessitating a theory of information that is not bound by the strict arrow of causal transmission.

geared towards disconnecting semantic information from traditional accounts anchored on correlation, covariation, or causation.

But first, we should probably turn to apparent weakness in covariational accounts. We have seen that Dretske's criterion of probability shifts all the way up to certitude (1) is probably too strong for a real-world application. So it is perhaps more charitable to take a more recent, nuanced stance as an example of correlational treatment of information. For instance, Shea's proposal (2018) stipulates that A's property of being F carries correlational information about B being G if and only if $P(Gb|Fa) \neq P(Gb)$. To elucidate the problem with such formulation and stress the difference between the similarity and the dominant correlational accounts, a doppelganger example is introduced (Miłkowski, 2020: 5). Let's think of a very secretive dictator whose appearance is shrouded in mystery. The foreign, hostile intelligence agency is willing to go to lengths to find out what the dictator looks like, but that is not possible. They do know, however, that the regime has a dictator's lookalike, a doppelganger that they use for certain occasions. Then if they manage to get the description (or, better, a picture) of the said doppelganger, they can reliably infer some properties of the actual dictator's physique. But here we should realize that the dictator's and doppelganger's appearances are statistically independent. Yet we are strongly inclined to say that one's appearance is highly informative of the other; the doppelganger's look carries information about the dictator's look. On traditional accounts, this kind of spurious statistical correlation should not afford valid inference. Yet, the hostile intelligence agency would still surely pay top dollar to perform this kind of surrogative inference (Swoyer, 1991).

This suggests that an alternative must be found in thinking about semantic information in terms of similarity judgments and structural mapping rather than statistical correlation. To formalize this, we turn to the correspondence-based

account of semantic information (Miłkowski, 2023), which builds upon Barwise and Seligman's (1997) pioneering work on information flow in distributed systems.

4.7.1 Infomorphism: The Strict Mapping

In this framework, semantic information relies on structural similarity between informational structures. Correspondence is formally analyzed using the concept of infomorphism. This relation is defined between two classifications. We understand a classification as holding between particular physical tokens and some type.

A useful intuition pump here is Optical Character Recognition (OCR) (see Figure 3). An OCR system involves classifications: an old typewriter produces physical ink marks (tokens) on a page. These marks introduce variations (smudges, tilts), but they are classified under a single type (e.g., the Latin letter 'a') by the typist.

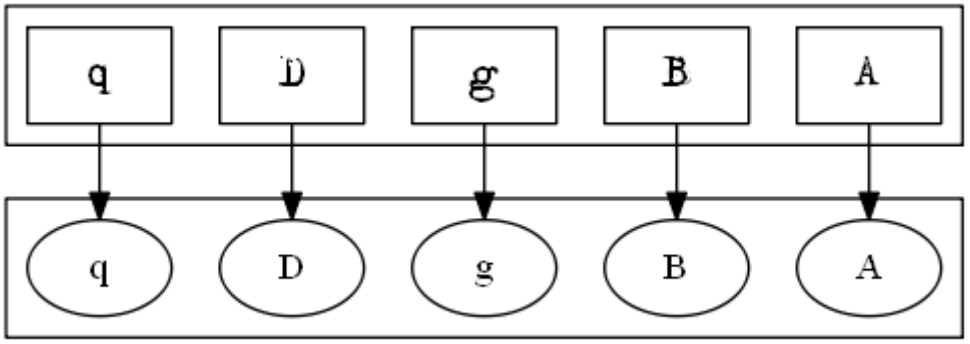

Figure 3: Typewritten characters correspond to Latin alphabet characters (adapted from (Mamak & Miłkowski, 2025)).

The OCR system extracts information because there is a correspondence between the images on the page and the digital Latin characters it outputs[12]. Formally, an

[12] One might worry that 'structure' is too vague a property to carry quantifiable information. However, Stephen Mann (2024) has recently demonstrated that the information carried by such structural representations can be rigorously decomposed. He shows that the relational structure (e.g., the ordering of elements) contributes measurable information (mutual information) distinct from and additive to the information carried by the components themselves. This formalizes the intuition that a structural correspondence (like a map or infomorphism) carries 'surplus' semantic

infomorphism exists when the type-token mappings are preserved between two classifications.

We can state this formally. Let two classifications be $A =< A, \Sigma_A, \models_A >$ and $B =< B, \Sigma_B, \models_B >$, where:

- A and B are sets of objects (tokens);
- Σ_A and Σ_B are sets of objects used for classification (types);
- $\models_A$ and $\models_B$ are binary relations between tokens and types that tell one which tokens are classified as being of which types.

An infomorphism f from A to B is a pair of functions $< f^\wedge, f^\vee >$ such that:

- $f^\wedge : \Sigma_A \to \Sigma_B$ (mapping types to types)
- $f^\vee : B \to A$ (mapping tokens to tokens)

Crucially, for all types $a \in \Sigma_A$ and for all tokens $b \in B$, the functions must satisfy the following biconditional:

$$f^\vee(b) \models_A a \iff b \models_B f^\wedge(a)$$

content that an unstructured list of features lacks. Mann's paper would require much more space to cover in order to do it justice. Thus, for pragmatic reasons, I do not pursue this avenue further here, but for now let it be noted that it is a promising line for future research within the IPP. A tentative proposal for such expansion comes briefly in the section 3.3 of ch. 3)

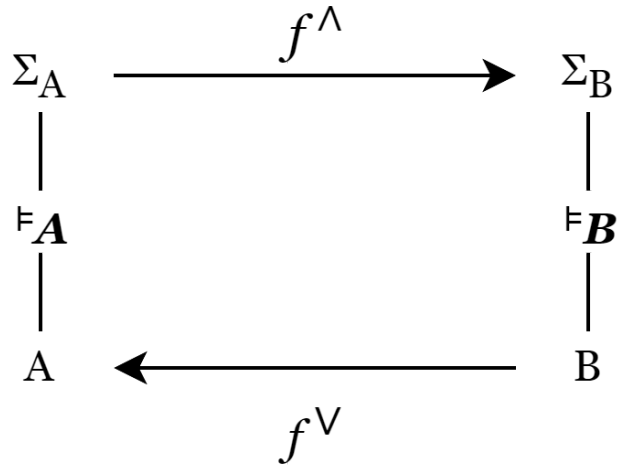

Figure 4: An infomorphism obtains between A and B; adapted from (ibidem)

Since Barwise and Seligman introduced this notion in the context of information channels, the functions $f^{\wedge}$ (read "f-up") and $f^{\vee}$ (read "f-down") can be intuitively understood as encoding and decoding. Thus, tokens of b can be encoded (via $f^{\wedge}$) and classified using the classification relation $\vDash_A$ as being of types a. When they can also be classified using $\vDash_B$ as being of types provided by the decoding relation ($f^{\vee}$), then an infomorphism obtains: there is a structure-preserving map.

4.7.2 Infocorrespondence: The Biological Reality

While the strict infomorphism models noiseless transfer, biological similarity admits degrees. The "Dictator" example above shows that even a perfect lookalike might not share all properties with the target. To account for this, we introduce fuzzy infomorphism, or infocorrespondence.

Infocorrespondence incorporates prototypical category membership and partial resemblance, inspired by fuzzy set theory (Zadeh, 1965). It can be defined for fuzzy classifications $A =< A, \Sigma_A, \vDash_A >$ and $B =< B, \Sigma_B, \vDash_B >$, where $\vDash_A$ and $\vDash_B$ are fuzzy binary relations between tokens and types.

An infocorrespondence f from A to B is a pair of functions $< f^{\wedge}, f^{\vee} >$ defined in the same manner, but the fundamental biconditional:

$$f^{\vee}(b) \models_A a \iff b \models_B f^{\wedge}(a)$$

holds for at least some tokens $b \in B$ and some types $a \in \Sigma_A$.

In fuzzy classifications, tokens belong to types or satisfy relations to a degree. This allows for approximate similarity rather than the strict structure preservation of an infomorphism. Quantifying the degrees of partial mapping allows one to measure the extent of correspondence. Crucially, this fuzzy correspondence still licenses surrogate reasoning: the system can infer properties of the distal target (the Dictator) from the proximal vehicle (the Doppelganger) because the structural mapping, while not perfect, is robust enough to preserve the relevant logic."

4.8 Israel & Perry's Situation-Theoretic Account

A sophisticated alternative to the standard views is found in David Israel and John Perry's (1990) situation-theoretic account. They argue that information is not an absolute property of a signal but is constraint-relative. Their central innovation lies in recognizing that information is grounded in real-world regularities, which they call "constraints," while simultaneously acknowledging that which information a given fact carries depends on the constraints to which an observer or system is attuned. I & P articulate several working principles that characterize their notion of information. First, informational content can concern remote things and situations, meaning that an indicating fact (such as smoke) can carry information about something not physically present in the fact itself (such as fire at a distance). Second, this content can be highly specific, with propositions about particular objects or states of affairs. Critically, however, indicating facts contain such information only relative to connecting facts or background constraints. The smoke

indicates fire only given the lawful regularity that connects smoke-producing processes with combustion. This constraint relativity means that the same indicating fact can carry different information relative to different constraint structures. Fourth, their framework acknowledges that many different facts, involving variations in objects, properties, relations, and spatiotemporal locations, can indicate one and the same informational content. As we will see later, This comes in handy, as such many-to-one mapping from vehicles to contents is perfectly acceptable within their framework[13].

This perspective shifts the focus from "decoding" a single absolute message to identifying which *constraints* a system is attuned to. The challenge, which we will return to in later chapters, is determining how a cognitive system selects or attunes itself to the specific constraints that allow it to exploit this information for adaptive behavior.

5. Putting the Desiderata to Work

Having reviewed the major theoretical candidates—Classical Shannon Theory, Algorithmic Information Theory (AIT), Semantic/Teleosemantic approaches, and the Ecological view—we must now play the judge and adjudicate them. We apply the eleven explanatory desiderata (D1–D11) established in Section 3 to determine which framework is most "friendly" to the cognitive science project.

Shannon's theory excels at Quantifiability (D1) and Uncertainty (D11), providing the gold standard for measuring bit-rates and uncertainty reduction. It also satisfies

---

[13] Importantly, though, it does not undermine injective mapping of contents within a certain informational pipeline. Constraints, in a way, are prior to establishing genuinely contentful informing and only limit the selection of possible coding schemes to choose from.

Universality (D2), as it applies to any channel regardless of the agent. However, it fails critically on Semanticity (D3) and Representationality (D4). As Shannon famously noted, "meaning is irrelevant" to the engineering problem. Consequently, it remains agnostic on Factivity (D6) and offers no account of Arbitrariness (D9) beyond arbitrary coding efficiency. Without supplementation, it describes the *capacity* for cognition, not cognition itself.

The ecological view offers a robust account of Objectivity (D7) and Structure (D9) (via lawful specification). It takes a hard stance on Factivity (D6): information cannot be false. However, this strength is its cognitive weakness. By defining information as lawful specification, it fails Semanticity (D3) (it has existence conditions, not truth conditions) and rejects Representationality (D4) entirely. It collapses Internality/Externality (D8) by denying the existence of internal vehicles "for" the system, making it ill-suited for explaining decoupled, offline cognition.

Dretske attempts to naturalize Semanticity (D3) via causal laws ($P(s \mid r) = 1$). This satisfies Objectivity (D7) but fails Factivity (D6) in a useful way: because it demands perfect correlation, it cannot easily account for misinformation (the Disjunction Problem) without retreating to a "learning history." It is also brittle regarding Universality (D2) and Portability, as channel noise destroys the informational relation entirely.

To foreshadow what follows: the updated IPP defended in this thesis (the *Neo*-IPP) will try to avoid the pitfalls of strategies on offer while supplementing their deficient aspects. To see them clearly, we suggest constructing a map/table for summing up the crosscheck of theories discussed against the desiderata we laid out. I suggest calling it an Explanatory Scorecard.

Desideratum	Shannon (Classical)	Ecological (Gibson)	Causal (Dretske)
D1: Quantifiability	**High** (Bits/Entropy)	**Low** (Qualitative)[14]	**Mixed** (Probabilistic)
D2: Universality	**High** (Agent-Agnostic)	**Limited** (Agent-Env Coupling)	**Brittle** (Channel Dependent)
D3: Semanticity	**None** (Syntactic only)[15]	**None** (Existential only)	**Attempted** (Indicator function)
D4: Representationality	**Agnostic**	**Rejected** (Direct Perception)	**Explicit**
D5: Leveling	**Abstract** (Computational)	**Flat** (One-level)	**Abstract**

[14] Gibson himself rejected Shannonian measures altogether. Nonetheless, as we saw (in section 4.4), there are now serious attempts to quantify Gibsonian information via measures such as 'affordance ratios', 'geometric invariants' or *tau* (τ), that is time-to-closure of an optical gap (e.g., Kim, Turvey & Carello, 1993).

[15] In traditional interpretation (and Shannon's himself), it's the renowned "semantic aspects of communication are irrelevant to the engineering problem" catchphrase (Shannon, 1948). This reading was recently defended by Lombardi et al. (2018). However, non-traditional accounts emerged (Issac, 2019), which we will analyze and utilize in more detail.

D6: Factivity	**N/A**	**Strict** (Cannot be false)	**Strict** (P=1)
D7: Objectivity	**Receiver-Relative**	**High** (Invariants)	**High** (Laws)
D8: Internality/Externality	**Collapsed**	**Externalist**	**Mixed**
D9: Arbitrariness	**Arbitrary Codes**	**Natural Mapping**	**Arbitrary**
D10:Epistemic Closure	**No**	**No**	**Yes** (Xerox Principle)
D11: Uncertainty	**Central** (Entropy)	**Rejected** (Specification)	**Eliminated** (P=1)

Table 1: The Explanatory Scorecard - table for cross-checking theories of information.

While the scorecard maps the three leading theories on offer, it does not automatically dissolve all conceptual ambiguities inherent in the term 'information. Neither it offers a solution yet. But to get a bit closer to more acceptable alternatives, we must now zoom in on three specific properties—veracity, receiver relativity, and the internal/external distinction—that demand a finer-grained analysis to be explanatorily useful.

6. Fundamental Properties of Information in Cognitive Systems

6.1 Veracity of Information

Time to confront an important issue that we have not yet touched upon, namely the question of whether informational states do have truth conditions. In other words, can information be veridical or falsidical? In the Shannonian theory, the answer is negative, as any kind of signal is as informative as its probability of being selected. The theory (at least in the orthodox interpretation) does not possess any semantic features that would allow for assessing the truth value of a given piece of information. Semantic theories, however, did not take this path and answer the question approvingly. More so, one of the primary reasons for constructing these theories was to introduce the property of truth-evaluability into information theory. For Dretske, the project of informational semantics was to achieve the ultimate goal of naturalization of knowledge, and he thought this end required the informational base to be truth-evaluable and truth-preserving.

It is important to remember that not only nomic theories, such as Dretske's, are committed to the alethic properties of information. For example, Cohen and Meskin (2006) defend a counterfactual account of information that is nonetheless truth-evaluable in the same sense as Dretske's conception. In their view, the spots mean measles because had measles not been the case, there would have been no spots. Still though, the information is carried if and only if the relevant scenario is actualized. They are both cases of win-or-go-fishing theories. So in both these theories, only truthful pieces of information count as information proper. After Piccinini & Scarantino we call this 'The Veridicality Thesis.' If the spots carry information that one has measles, it follows that one has measles. It states that if a system possesses information about p, that entails that p obtains. According to semantic theories, speaking strictly, false information simply does not exist.

That is surely against the common usage of the term and folk understanding of information. After all, we hear about false (or, recently, fake) information all the time. We debate how to combat it, flag it, and filter it out effectively. What about the case if one does not have measles and the spots just accidentally look really similar to the ones that are caused by measles? According to the classical theory of semantic information, that would constitute the case of misinformation. The ability to be truth-evaluable was precisely what motivated some theorists of semantic information to differentiate natural from non-natural information. Recall Grice, who posited that the difference between natural and conventional signals hinged just on the fact of whether signals could be true or false or not. The non-natural conventional signals did not come in a package with entailment of truth, whereas natural signals in fact did. Smoke cannot carry information about fire if the fire is not there, but the bell's ringing on the bus could carry false information about whether the bus is in fact full or not. This inability to account for failed relations was often used as a charge against informational theories of representation, as the hallmark property of representing is the possibility of misrepresentation. If informational states are by definition true, they seem not to be able to account for invalid representation.

But according to some philosophers (Shea, 2007, Millikan, 2000, 2004; Piccinini & Scarantino, 2010), there is no good ground to say that the sentence 'the spots carry information about measles, but one does not have measles' makes no sense. If that is the case, we should reject or modify the Veridicality Thesis. What then would the aforementioned sentence mean—or, better for our purposes, what kind of natural information is carried by such a signal? The response is to state that the signal S carries information p when the fact that p obtains is more likely to happen when the S is present. Piccinini and Scarantino call this proposal 'Probability Raising Thesis for Natural Information (PRT(n))':

(PRT(n)) If a signal s being F carries natural information about an object o being G, then $P(oisG|sisF) > P(oisG| \sim (sisF))$.

On this view, a signal s in state F can carry natural information about an object o being G simply by raising the probability that o is G, whether or not o is in fact G (Piccinini & Scarantino, 2010:317). The rejection of the Veridicality Thesis, however, does not imply that an agent cannot possess natural information. Contrary to VT but consistently with PRT(n), a signal s being in state F can carry natural information about an object o being G even though o is not G. It follows that an agent A can have natural information about o being G whether or not o is G, provided a reliable correlation exists between the signal type received by A and the event type instantiated by o being G (Piccinini & Scarantino, 2010:319). The gist of this approach is to understand information carriage in terms of reliable correlation. Reliable correlations between the variables in the environment (or the internal states) are such that their receivers can count on them to systematically (more often than not) be the case in the case of future (and counterfactual) instances. This approach of reliable correlation as sufficient grounds for information, as we will see later on, has been wildly influential in interpreting neural signaling and is oftentimes the (tacitly or explicitly) assumed and implemented understanding of what it means for a signal to carry information about something (e.g., Borst & Theunissen, 1999; Meister & Berry, 1999).

The structural similarity account, on the other hand, takes a step back from truth conditions altogether, offering instead just satisfaction conditions. Satisfaction conditions may, but need not, provide grounds for truth-evaluability. For declarative pieces of information, satisfaction conditions are indeed the alethic evaluation: whether it is true or not. But for instructional pieces of information, we want to evaluate them regarding whether they have been followed or not (successfully executed or not). Both these operations take binary values but are not

equal. For the former case, the declarative piece of information is true when there exists a corresponding relation between the given piece and the state it is supposed to inform about. This satisfaction condition is analogous to Tarskian classical theory of truth (1933), but here it is spelled out in terms of the aforementioned relation of structural similarity. The semantic condition for similarity is then defined:

'(SEM-TRUE) If infomorphism obtains between A and B, then A is true about B and B is true about A. In other words, the satisfaction conditions of A about B are the same as the conditions under which the infomorphism between both holds. The truth-value-truth-is ascribed to a classification A depending on whether the infomorphism between A and B holds.' (Miłkowski, 2010:13)

But as we have already discussed, because the similarity is a graded property, such strict criteria for informational relations are problematic, so the amended definition can be adequately restated in terms of loosened relation of infocorrespondence. The updated definition goes as follows:

(SEM-TRUTH-DEGREE) The truth-value of an information vehicle classified as A that stands in an infocorrespondence C with classification B is determined by the degree of similarity with which C holds. (ibidem)

What is interesting, though, is that Miłkowski's account is agnostic about the possibility of defining falsity somewhat symmetrically to the presented definition of SEM-TRUE or SEM-TRUE-DEGREE. At least three possibilities are open to account for falsity. First is to state that A is false about B if and only if there is no infocorrespondence between them. Second is to state that where there is no relation of infocorrespondence between the two, there one is simply not about the other, and the problem of misinforming (or mistargeting) does not concern the general theory of semantic information. The third option is to say that

misinforming can only be accounted for on pragmatic grounds, where some additional conditions where the normally informational relation fails to inform due to abnormal conditions (an analogous strategy is chosen for accounting for misinforming based on his theory of information by Floridi [2008]).

6.2 Objectivity of Information and the Problem of Receiver Relativity

We have so far discussed various accounts and aspects of information, but we have hardly touched upon yet another massively important question of the objectivity of information. We mentioned briefly that one of the sources (sic!) of excitement for the Shannonian theory and its wide adoption was the hope that it would provide common communicative currency for all sorts of agents. If we had this kind of currency, we could then try to cash out notoriously slippery terms known from psychology and philosophy, such as propositions, beliefs, or knowledge, in terms of this common currency, escaping the muddy waters of subjectivism and qualitative description (Adams, 2003). This promise still lingers, as we have seen at the beginning of the chapter, where BCIs and neural decoding hypotheses were summoned.

It seems reasonable, then, to claim that the entire project assumes quite explicitly that information is objectively 'out there.' This is particularly visible in neuroscience, where, regardless of any philosophical niceties, researchers shoulder through them, reifying the 'information flow' within the neural circuits. I suggest calling this tendency a 'hydraulic fallacy,' where information is haphazardly conceptualized as liquid-like matter that flows through designated pipelines (pathways). This kind of thinking is motivated by research programs such as connectomics (Seung, 2012) and other structural analyses of the topography of neural circuits. Mapping the connections and measuring the flow—however hugely valuable to understanding neural processing (which I do not deny by any means

here)—is not sufficient without the theory of semantic information within the said mazy pipelines.

That might strike you as a platitude, but suspiciously numerous studies in neuroscience seem not to care too much about being careful in stating what is really mapped and measured in neural information research (Fresco, 2010; Fresco, 2014). To give a flavor of this imprecise usage, let's bring up Meister and Berry's influential account of research on the retinal processing of visual information. In their 1999 paper with the symptomatic title 'The Neural Code of the Retina,' they do not hold any horses:

> 'Finally, the analysis of retinal processing will benefit greatly from the fruitful application of information theory. It has long been recognized that the essential substance transmitted by neurons is not electric charge or neurochemicals, but information (MacKay and McCulloch, 1952). In analyzing a neural system, it is essential to measure and track the flow of this substance, just as in studies of the vascular system one might want to measure blood flow. Fortunately, there exists a well established formalism by which to measure information and analyze its processing [Shannon and Weaver, Rieke et al., 1997]' (Meister & Berry, 1999: 447).

That seems prima facie as a hasty, conflated conceptualization. That might be too charitable, though. In fact, it is a consistent strategy of awkward reification that pervades not only the entire paper but numerous studies in the field. This 'awkward reification' is powerfully and poignantly criticized by Rathkopf in terms of informational objectivity and receiver-relativity (Rathkopf, 2017, Rathkopf, 2018). Rathkopf points at the 'regular' mechanistic accounts of flow, such as the classical Harvey model of blood flow in the human body. In that case, there is nothing in principle that would block the possibility of extracting the flow from within the

system without the flow still occurring with roughly the same function. This possibility is not only theoretical now but actually quite casually performed now during the extracorporeal membrane oxygenation (ECMO) procedure. During this (quite complicated technically, obviously, but already standard) procedure, the blood is taken out of the ailing body for the function of pumping, oxygenation, and deoxygenation of blood to be done externally in a dedicated piece of equipment, sometimes called the lung/heart machine. If information were like blood, it would be an intrinsic substance (a "commodity") that retains its identity regardless of where it is. It would be universally moveable, or, as we will use this term, portable. In very plain terms, one could take a "thought" from one brain, put it in a bucket (or another brain), and it would still be that thought.

Rathkopf poignantly remarks that any kind of neural information flow cannot be ‘taken out,’ as in the blood flow example. So the metaphor of neural information flow collapses. In other words, while the analogy is intuitive, it risks a reification fallacy. Blood is a physical substance that retains its identity regardless of the vessel. Information, by contrast, is a relational property that depends on the statistical structure of the receiver (the 'consumer').

The reason for the unfeasibility of the hypothetical ‘take-out’ is precisely the problem of receiver relativity. An information system is receiver-relative just in case the mutual information between the set of messages and the set of signals depends on facts about how the underlying correlation is exploited by downstream mechanisms (Rathkopf, 2017: 6). The crucial point here is that the nature of Y’s function depends on the manner in which its activities are exploited by downstream mechanisms. Recapitulating, the gist of the argument is that we seem to be unable to unanimously ascertain the ‘real’ or ‘best’ or ‘actual’ measure for a given correlation. Researchers probing the neural structures in this way are always entangled by the experimental setup and subsequently defined ‘purpose’ of given

informational signaling. This remark resonates well with Rosa Cao's stance, who criticizes accepting the promise of teleosemantic accounts of neural content to be free of the problem of interpretation, pointing out that neural semantics is equally prone to Quinean critique of indeterminacy as 'regular' semantics.

A striking empirical demonstration of this tension—and its resolution—is found in the "Brain-to-Brain Interface" (BTBI) study by Pais-Vieira et al. (2013). In this experiment, an "Encoder" rat in Brazil performed a tactile discrimination task while its motor cortical activity was recorded and transmitted via the internet to a "Decoder" rat in the United States. The decoder received this information as intracortical microstimulation (ICMS) in its somatosensory cortex and successfully utilized it to select the correct lever, despite having no direct sensory access to the task cues. This result compels us to adopt a sort of conditionalized receiver relativity. Indeed, the experiment demonstrates that neural information is indeed portable (capable of being detached, digitized, and transmitted across a channel), thereby refuting a radical localism. However, the successful transfer was contingent upon the 'Decoder' rat being strictly conditioned (trained) in the regimented fashion to interpret the specific stimulation pattern (to bind them with wanted behavior). Thus, neural information is not a universal fluid like blood but a conditionally portable vehicle that requires a synchronized consumer architecture to retain its semantic content.

This distinction between the vehicle moving *between* systems and the content being realized *by* the consumer leads us directly to the next crucial disambiguation: distinguishing the information available *to* the system from the information observed *about* it.

6.3 Within Does Not Equal About: A-Information vs. F-Information

This issue points to a crucial theoretical distinction that must be made to avoid the hydraulic fallacy. We must distinguish between **information-about-the-system (*a-information*** in terminology I will use**)** and **information-within-the-system** (**information-for-the-system or *f-information***). “Information-about” refers to the statistical correlations and mutual information measures that a scientist (the external observer) extracts from neural data—information-as-measure. "Information-within" refers to the actual instructional content or signals that the system itself utilizes to guide its behavior—information-as-substrate. The hydraulic fallacy occurs when researchers mistake the former for the latter, assuming that because *they* can measure a correlation (about the stimulus), the brain must be using that correlation as a vehicle (for the stimulus).

This philosophical problem manifests methodologically as the Decoder’s Dilemma (Ritchie et al.,2019; Rathkopf, Heinrichs & Heinrichs, 2023). Neural decoding algorithms can extract information from brain activity with high accuracy, but this leads to a conflation. As Rathkopf et al. (2023) argue, successful decoding demonstrates only that information is present in the signal (correlation), not that it is functionally relevant (causation) or used by the organism. If information is receiver-relative, and we do not know *who* the receiver is (the downstream neuron? The whole organism? The homunculus?), then our decoding efforts may be identifying artifacts of our statistical power rather than the actual "mechanics of thought." Without solving this dilemma—without establishing when information-about becomes information-for—the IPP remains vulnerable to the charge of reification.

So, if the objective status of information processing is in danger, does one gain anything by adopting the informational framework, really? Did we not throw the

baby out with the water, accepting receiver relativity and hence partial 'mind-dependence'? Not necessarily. One tentative but promising response is offered by Fresco (2021), who argues that information can play an important explanatory role in the mind and brain sciences, in accord with cognitive sciences' original desiderata. What do you think? These desiderata can be conjointly labelled 'the cognitivist principles.' Cognitivism, very roughly sketched here, is a view that assumes that cognition involves computations over internal representations that happen within a neural system (brain or its counterpart). As such, cognitivism is quite a crude and amorphic position but was conceptualized by the 4E cognition theorists, who wish to dismantle it. Nevertheless, it seems reasonably uncontroversial (also to these criticisms) that cognitivism as such is very much still a mainstream view—both in philosophy and in mind and brain sciences (perhaps even more so in the latter). As a representative illustration of this view and its 'cognitivist principles,' let's use Rowland's (2011: 110-111) criteria for what constitutes a cognitive process.

P is a cognitive process if:

(1) P entails the processing of information

(2) This information processing has the proper function of making available information that was previously unavailable (at a personal or sub-personal level).

(3) This information is made available by way of the production, in the subject of P, of a representational state.

As we can see, especially inferring from the second premise, the core property of cognitive activity is, in a sense, extracting and processing new information. This information does not necessarily have to come from the external environment, but even when applied to interoception, it has to bring about (make available) new,

previously unavailable information. What is also important to note is that information is then a resource for the cognizer to mine from the information-rich environment. Information is a commodity, inasmuch as it enables different cognitive agents to utilize it in a sort of neutral way. Dretske writes, 'Information [... is] an objective commodity, something whose generation, transmission, and reception do not require or in any way presuppose interpretive processes' (1981: 8). It is only later when this raw material is processed or encoded with the minds. It is worth noting that the importance of such conceptualization for cognitivism is conceded even by its contestants (Hutto and Myin, 2018, Menary, 2013).

We can now understand why the stakes for defending the mind-independence status of information are so high and the problem is so crucial to the cognitivist debate and, as such, to the very core principles that are openly or tacitly accepted within the cognitive sciences. If there is no agreement on the notion of information, then there is no congruent research program on explaining cognition, and the entire cognitivist paradigm is in grave danger. According to Fresco's analysis, however, this worry is too hasty. According to him, even receiver-relative information can be regarded as epistemically or scientifically objective as long as it is shareable and evaluable by different receivers, which would fall within the bounds demarcated for Popperian 'intersubjective agreement' (Popper, 1979; as it is commonly accepted and used in science. Even if Fresco is right in his Popperian position, there is ample reason to examine and clarify the practical usage of information talk and its logical consequence for explanatory powers of cognitivist theories. The non-monolithic theory can be equally workable, but it is much more complex to adequately analyze and describe.

7. Conclusions

Based on what we've discussed so far, it seems clear that two general strategies for explaining the convoluted informational flow are possible. First, attempts at providing an overarching theory (or the mirage of it, hiding real, sometimes contradictory assumptions) of all usages and contexts of information. Usually some version of Shannonian theory is used for this purpose. Unfortunately, though, usually it turns out that under close inspection these accounts are either overextending the scope of Shannonian theory or making unsubstantiated claims about its applicability to various, disconnected domains. A second approach is to provide a taxonomy of uses of information and clearly delineate sub-meanings from one another, labeling them accordingly and providing clear criteria for ascribing, measuring, and manipulating a particular subtype of information. It is the second strategy that we want to employ and follow more clearly in the following chapter, where the aim will be to look more closely at the variety of information strictly in the neural signaling context.

We will start with an analysis of a range of proposals by both philosophers Sprevak (2020), Cao (2018, 2020), Rathkopf (2020), Orlandi (2014) and scientists Haxby et al. (1991), Beer & Williams (2010), Kriegeskorte et al. (2004) to confront this problem in a regimented fashion. What will be especially interesting to us will be how to map theoretical categories, such as senders/transmitters/receivers, but also noise (that we have not discussed much so far) of the signal. Two key relationships that any sufficient informational theory should be able to account for are information-representation and information-content dyads (or give reasons to explain them away, as in the case of deflationary accounts). But what is equally important and perhaps less discussed in literature are the questions of information vehicles and their format. We will also be able to better grasp what is meant by positing operations on information, such as consolidation, integration, suppression,

amplification, etc. Only then will we be ready to analyze and assess the neural coding theory and its critics (Nizami, 2019, Bickhard and Terveen, Brette, 2019. Zahnoun, 2025) more soundly.

Chapter 2: Representation

1. Can't Get Me Out of Your Head (Indispensability of Representations)

1.1 Why Representations Anyway?

Firstly, it is reasonable to say a few words about why we should be concerned with representations in the first place. Admittedly, the representational theory of mind is so widely accepted in cognitive science that it might appear not to need any further justification. Since the advent of cognitive sciences in the 1950s, the study of cognition became increasingly connected with the rapidly expanding computer science. During this so-called 'cognitive revolution,' computation established itself as a virtually universally accepted procedure involving operations on symbols. Symbols were considered representations, so computations were clearly defined over representational states. 'No computation without representation' hailed the famous motto coined by Fodor (1975). The good old-fashioned artificial intelligence paradigm (GOFAI) received heavy blows during the AI winters of the mid-1970s and late 1980s and lost traction, largely due to the rediscovery of connectionism and reinvigoration of ecological and interactivist movements. Nonetheless, the princely status of representations as main theoretical posits of cognitive science remained almost unscathed for much longer and remains so.

Why are representations so enticing for cognitive theorists, then? Historical reasons and continuity of research traditions, vocabularies, and paradigms are surely important, but certainly not sufficient to explain how pertinent the notion turned out to be. The received view contends that there is special explanatory

purchase that representations provide for making sense of how cognitive processes operate. How they appear and why they are so successful. In somewhat crude terms, researchers for decades were certain that it is simply impossible to understand what happens inside the heads of people (in the case of psychology) or other innards of cognitive systems (in disciplines under the artificial intelligence umbrella, more broadly) with no appeal to representations.

Let's turn to a classical example often summoned to demonstrate this mechanism. Roger Shepard and Jacqueline Metzler's (1971) work on mental rotations was pioneering research on the phenomenon. The problem in question concerned the ability of subjects to successfully ascertain whether two 3D objects were indeed the same object but presented in different angles (planes of rotation) or simply different objects. Shepard and Metzler measured the reaction times of participants' answers and found that they differed systematically; namely, there was a linear relation of time increase to the level of rotation. The farther the compared object was from the original (baseline) position, the more time it took the subjects to assess whether two objects were in fact identical or not.

To explain this phenomenon, researchers thought it was necessary to invoke a theoretical posit—a mental object that was being manipulated in the participants' minds. What is important to note is that the object was purely theoretical, insofar as it could be neither observed nor directly manipulated by the scientists. They also did not aspire to do so. Representations were mobilized solely to explain the experimental results. Shepard and Metzler thought (and most psychologists agreed) that it was the only avenue for accounting for the phenomena obtained experimentally.

1.2 The Imagery Debate

It is important to acknowledge, however, that this interpretation was not uncontested. The field was soon engulfed in the "Imagery Debate," pitting Stephen Kosslyn, who argued that mental images function as quasi-pictorial representations that preserve spatial structure, against Zenon Pylyshyn, who contended that the underlying format of these representations was propositional (language-like) and that the phenomenology of imagery was epiphenomenal (Pylyshyn, 1973; Kosslyn, 1980). Pylyshyn argued that the linear reaction times could be explained by tacit knowledge or the structure of the underlying propositions, without positing a rotatable internal "picture." For a time, Pylyshyn's critique led many to believe that pictorial representations were theoretically suspect. Nevertheless, despite the fierce debate over the *format* (image vs. proposition), both sides agreed on the necessity of *representation*. Whether the mind rotates a picture or processes a description, it is trafficking in internal states that stand in for external objects.

Mental rotation is a tiny fraction of psychological research, obviously, so the theoretical landscape would not necessarily be representative of the entire field. But it became a paradigmatic case of theoretical explanations in both psychological literature and in philosophy of science. Representations are mental objects that retain some important information about the environment and/or the task to be achieved and are key to both veridical perception and effective action. It makes intuitive sense—we correctly identify the branch hurled at us by an angry tempest during the autumn walk and successfully evade it because we have an accurate representation of the branch and its position (and also the future colliding trajectories of ourselves and the branch). It is painfully more obvious in the case we fail in our representational endeavor. One gets hit because the representation one had in mind was a poor one.

1.3 The Representational Job

Before we tackle the case of what representations are supposed to be, there is yet another thread of the ‘why’ question: what kind of job are they supposed to perform? The logic goes roughly as follows: when we clearly understand what these theoretical objects are supposed to do, then we will get a better grasp of what they are. To use a parable, watching a documentary about the content moderation on social media and learning first about Facebook policies, goals, values, and criteria employed, I was able to much better understand what kind of people are required to do this job. One realizes how tenacious, mentally strong, and decisive the virtual policemen have to be, and that raises the understanding level even before they are introduced in person in the movie.

This kind of strategy is often employed by theorists when talking about the ‘representational job’ to explain the nature of representations. A few influential accounts were formulated. For example, Haugeland (1991) outlines three main representational duties:

1. The system coordinates its behavior with the environmental features even when these features are not reliably present.
2. It copes with such cases by having something else "stand in" and guide behavior instead of immediate stimulation.
3. The "standing in" is systematic and part of a more general representational scheme of related representational states.

The ‘standing in’ relation is crucial and almost universally accepted as a primary feature of representation. It is somewhat unsurprising. Regardless of the particularities of representing and the represented, one ‘shadows’ for another, doing its job in lieu of its target. Representations are emissaries, even if we do not agree on who communicates with whom and what kind of messages are delivered

or even in which language or format. This ‘standing-in’ relation will form a most rudimentary, minimal form of account of representation throughout the literature.

1.4 Critique of "Standing-In"

However, we must be cautious with this metaphor. Aaron Sloman (2011) offers a critique of the notion that representations only "stand in" for their targets. He argues that such a "standing in" typically involves taking the place of something in a specific role, like a substitute teacher standing in for a regular teacher. But mental information structures rarely function this way. You can eat a banana, but you cannot eat the *information* about a banana. The use of information structures is fundamentally different from the use of the things they represent. Manipulating a recipe for a cake produces different results than manipulating the cake itself. Thus, while the "standing in" metaphor captures the idea of surrogacy, it must be understood as enabling specific *kinds* of operations (planning, reasoning) that the original object does not support, rather than simply replacing the object in a causal chain.

Yet another proposal, perhaps even more reverberant, was Ramsey’s 2007 ‘job description challenge.’ Ramsey also agrees on the importance of standing-in relation. However, he thinks that in the case of elucidating the job, there is an important distinction to be made. For the ‘job description challenge’ covers only one aspect of explaining what actual representation is supposed to be. This is a functional aspect, which in his words equates to establishing a "set of conditions that make it the case that something is functioning as a representational state. In other words, it is the set of relations or properties that bestow upon some structure the role of representing" (Ramsey, 2016:5). We should differentiate, though, the functional aspect from the ‘what’ question, namely establishing a separate set of

relations (or features), ones that „bestow upon some representational structure its specific representational content" (Ramsey, 2016:4).

Ramsey thinks that the former is very important for explaining when we can declare that some object clears the bar for genuine representation. In his view, in line with Haugeland's position, representation is first and foremost warranted in cases where representational structures' role is set within the functioning of a wider cognitive system. In other words, the content of representation itself is not enough, as long as we do not know what kind of job it performs within a bigger whole. We set the 'what question' aside for later, when we will discuss the main properties of what representations are later on. Now we move forward with the functional.

2. The Dialectic: Challenges to Representation

2.1 The Hunger for Representations

Representations were thought to be necessary, particularly to perform more sophisticated, complex, or advanced tasks. Simple, rigid, or automatic processing, such as unconditional conditioning, reflexes, or fixed behavior, could in theory be performed more minimally. The robot in a maze can make the traverse successfully using the 'dumbest' strategy of following pre-fixed turns at given junctions, provided it was designed for this particular setup. But almost any slight change renders him helpless and unable to correct his strategy. But if the environment is dynamic, randomized, or changeable, a more sophisticated mode of interaction is needed. The more complex a situation becomes, the more useful it is to get access to information about the environment. Still though, mazes are structured enough that a clever pathfinding algorithm or simple probing might do the trick.

Although we can imagine more complex scenarios when using past or future-oriented information is needed to speed up or even enable completion of the task. Even in the case of a maze, if the states of the robot are reset at every try, one that has access to previous successful attempts (memory) would be much better off than the one starting from scratch. But in the open world, sometimes even more sophisticated techniques are employed—involving imagination, internal simulations, multi-stage recall (piecing together information from different situations or timescales), planning, etc. These are the kinds of tasks that are extremely difficult to imagine without utilizing representations.

Clark and Toribio, in a classic paper (1994), dubbed these kinds of problems the 'representation-hungry' problems. Ones that are hard enough that they seem to be calling for representations to be employed. Planning, imagining, and counterfactual reasoning are cognitive acts in which representations have to be mediated to make them possible. Otherwise, how would we be able to think about things that are not immediately available to us during the 'cognitive moment' (or short timescale), so to speak?

Thus, representation-hungry problems are mainly of two strands:

1. Cases that involve reasoning about absent, nonexistent, or counterfactual states of affairs.
2. Cases that involve selective sensitivity to states of affairs whose physical manifestations are complex and unruly (e.g., the ability to pick out all the valuable items in a room) [Smortchkova et al.,2020:9].

This matter was pressed further by Andy Clark in his (1997), where he defined representation-hungry problems by contrasting them with the ones not requiring representational modulation. He characterizes the latter as situations "where the inner and the outer exhibit this kind of continuous, mutually modulatory,

non-decoupleable coevolution, the tools of information-decomposition are at their weakest (emphasis mine). What matters in such cases are the real, temporally rich properties of the ongoing exchange between organism and environment".

In a sense, what Clark is trying to say is that there are cases in which the information readily available in the environment is sufficient; thus, no representational support is needed for a cognizer to succeed. What I want to stress here (because it will be important for the purposes of this thesis) and what is perhaps not the point of focus for other commentators (Kieverstein & Reitveld, 2018; Degenaar & Myin, 2014) is the remark on information-decomposition. According to Clark, the information-theoretic description works most poorly in the cases that are not representation-hungry. For simpler scenarios, the dynamic picture is enough to explain what is going on; only when things get complicated is there a need for cavalry to arrive.

2.2 The Anti-Representationalist Attack

Clark and Toribio were not anti-representationalists, though. Or at least they do not form the center of the anti-representationalist camp. Their analyses and differentiation between representation-hungry problems and ones that do without were intended to show that the latter are much more common than we think. Thus, the theories that emphasize the potency of non-representational, non-informational explanations are much more convincing in their view. We can say that their position was the hallmark of advancing anti-representationalism.

Anti-representationalism was in many senses a truly revolutionary movement. We have already said that representationalism was so widely accepted in the cognitive sciences, assuming even that it became kind of a zero point of any further theorizing about cognition. A conception so entrenched that it threatened to make what was originally a theoretical posit into something more similar to an axiom (not

in a strictly formal sense, of course). Degenaar and Myin (2014:3640) go as far as saying that ‘indeed, representationalism has become so ingrained that representational assumptions have acquired an axiomatic status, as if "representational attributions are not the result of, but the prerequisite for, theoretical development. Representations are invoked even before the theory starts" (Tonneau, 2012:330)’. It is quite understandable why it took some time for challengers to the mainstream to appear on the scene and get vocal enough to be heard.

Admittedly, some forms of anti-representationalism were present forcefully in philosophy proper. Most famously, Rorty’s (1979) critique of epistemological tradition claiming that knowledge is a reflection of (and not only ‘on’) was dubbed anti-representationalism (Salis, 2023). This line of thought was continued by Brandom (2013) and others in the Sellarsian tradition of philosophy of mind. But this argumentation was mostly directed at classical philosophical positions (mostly Kantianism) and was not directly applicable to the problem of representation in a more technical sense as used in cognitive science as the theoretical posits within science, as we discussed them.

Anti-representationalism proper (for our purposes at least) gained traction in the 1990s because of dynamical theories (van Gelder, 1995) and reinvigorated connectionism (Dayan & Abbott, 1995). Philosophical ideas of Clark also were influential contributors to the movement of undermining the representationalist fortress. By 2000, Anthony Chemero was bold enough to declare that ‘natural cognitive systems do not traffic in representations’ (Chemero, 2000:627). But antirepresentationalism, as most philosophical positions perhaps, comes in shades. There are different levels of radicalism among theorists who are suspicious about the reign of representational explanations. These doubts range on at least three axes in my opinion—range, makeup, and realism.

There is disagreement about the ontic status of representations—how real they are. Going last to first, in realism debates, skepticism ranges from mild instrumentalism (Colombo and Seriès, 2012, Sprevak, 2013) through 'useful but false approximations' (fictionalism—Toon, 2023) to flat rejection (eliminativism—e.g. Chemero, 2000). By *makeup*, I understand the propensity to 'trim' the notion of representation, or in other words, the level of commitment of the term. Old-fashioned understanding of representations is indebted to Fodor and GOFAI, as full-blown language-like structures[16] are, in this sense, more ontologically committing than, for instance, 'lighter' teleological theories. On the other side of the spectrum, there are various 'deflationary' or 'minimal' theories that still declare allegiance to representationalism but claim that the term is to be understood more stringently. Finally, there is disagreement on how much of cognitive abilities are mediated by representations.

Clark and Toribio's work was the attempt at this kind of renegotiation of territory. According to them, not all, but arguably more than previously thought, can be wrested from the representational grasp. Today, radical enactivist movements (e.g., Myin and Hutto, 2017) go all the way, claiming that there is simply no place for representations whatsoever, and as everything can be explained dynamically, the king with no domain should be expelled for good. Now I want to briefly overview these varieties of anti-representationalism to accentuate the differences more poignantly.

Eliminativism about representation became quite a diverse movement. United under the common catchphrase of 4E cognition (embodied, embedded, enacted, extended; Rowlands, 2010), anti-representationalists deny that representationalism

---

[16] It is worth mentioning that at the time, the connectionist movement (Rumelhart et al.,1986), which operated at the sub-symbolic level of node activation, considered their 'currency' to be representational as well.

is still the only game in town (Newen et al.,2018). According to 4E theorists, computationalism (which usually hinges on representationalism) is but a fad stemming from the great leaps in the powers and achievements of actual computers. It seems naive, according to them, to jump on yet another bandwagon of the currently new shiny technological paradigm for explaining the mind.

The study of memory is a great example of that, as demonstrated by Draaisma (2000). Paradigmatic examples of memory-like mechanisms changed in the 19th century as soon as the new gadget appeared on the scene - camera obscura, daguerreotypes, photography, and finally motion pictures were all lauded at their heyday as models for understanding how human memory works. All seem laughably naive in hindsight. Why then should we copy the same mistake, 4E theorists seem to ask in a sort of pessimistic, metainductive way.

Eliminativism is their answer to the problem. As we remember, computationalism in its traditional form assumes that cognitive processing starts with transduction of energy (coming from the environment in the form of stimuli) into symbolic representations. Computationalists may disagree about the nature of this process, how it happens, and particularly on which stage of cognitive processing (how far upstream). But virtually all agree that it happens according to some rule-based processes. As an effect, representations are formed. Internal transformations of these representations according to some other set of rules bring about cognitive phenomena - speech generation, problem solution, perception, etc.—as an output.

According to eliminativists, stories about mental algorithms operating over internal representations are, well, stories indeed that researchers used to tell themselves in lieu of the better theoretical apparatus. Instead, 4E cognition draws from different sources. Maturana and Varela's Santiago Program of Cognition introducing autopoiesis is one of them (Maturana & Varela, 2012). Even though the former was

hesitant about extending his autopoietic ideas about the ability of living systems of self-generation outside the (originally considered) molecular domain, the former continued with extensions. His 1991 book *The Embodied Mind,* written with Thompson and Rosch (1991), was an influential source for 4E theorists, especially as its anti-internalist, body-centric ideas were in line with Clark and Chalmers's (1998) *The Extended Mind*. Flor and Hutchins (1991) were perhaps the first to introduce the idea of distributed cognition so openly. It shifted the unit of analysis from individual agents into bigger ensembles - cooperation between agents and agent interactions with artifacts (cognitive and material) as a baseline for cognitive systems.

But perhaps the most influential of all was not the most recent of tradition - the ecological perception of J.J. Gibson we discussed in section 4.4 of ch. 1. And the Gibsonian paradigm, as put forcefully by Shapiro, 'rejected nearly every plank of the information processing model of cognition that computational cognitive science epitomizes' (2021, section 1). Ecological psychology flatly rejected the claim of representational theorists that the environment is in a sense not enough to perform successful cognition. In their words, stimuli (as messages from the outside) were 'impoverished' (Chomsky, 1980).

The gist of the argument was concentrated on known cases in perceptual science where solely the perceptual input seemed to be insufficient for forming a ready percept. A so-called 'inverse problem' in vision science was a prime example of that. In short, an inverse problem arises (and it has been known since Helmholtz's times) when just the visual cues cannot differentiate between two percepts that are in line with what is mirrored in the retina. Stemming from laws of optics, for instance, different rectangles when projected onto the eye's surface are not recognizable (there is many-to-one mapping). The same happens with the size constancy phenomenon—same-sized (in absolute terms) objects appear

perceptually as differently sized when they are presented out of perspectival relations (with different depths), even though if one were to cut one out and overlap it over another, they would appear perfectly identical.

This led to the notion of 'poverty of the stimuli,' that external sources of information are not reliable enough for an internal processor to resolve potential conflicts in interpretation. Internal processing was needed to add the relevant context to perform this operation reliably and accurately. Some information, in an important sense, was thought to be necessarily coming from inside. Computationalists then tacitly (but sometimes openly - as e.g., Rock, 1983 or Fodor & Pylyshyn, 1999) assumed that perception is in part an active, interpretative process. Usually this process was supposed to be inferential-like, following the Helmholtzian tradition of unconscious inferences (Helmholtz 1867).

Ecologists repudiated these beliefs on a fundamental level. They claimed that the perceptual system is not limited to just the system usually thought to comprise it (roughly the sensory apparatus + nervous system). According to them, perception is an active process in which the environment is not a passive background but an essential part. Active explorations and dynamical coupling between the rich environmental features sending out not information, but affordances (not just messages, but offers to interact) - only this forms a more complete picture of the perceptual system.

As such, ecological perception is readily understandable as the perfect jumpstart for 4E's attack on the mainstream's "contingent intracranialism" (Adams and Aizawa, 2008), where cognition happens only inside the cognizer's head and is strictly instantiated by brain processes. Antirepresentationalism is then a sort of anti-Cartesianism. This renouncement of internalism goes far, as 4E theorists have recently argued that even paradigmatically representational features of

modality-general cognition, such as speech production (e.g., Glenberg and Kaschak, 2002) and memory (Casasanto and Dijkstra, 2010), can be accounted for on anti-representational grounds.

However, we must be careful not to conflate all critics of the classical IPP into a single eliminativist camp. While Clark's work on the "extended mind" and "representation-hungry" problems provided ammunition for the anti-representationalist turn, Clark himself does not reject representation. Rather, he advocates for a sophisticated, action-oriented form of representationalism (Hutto, 2013). For Clark, the recognition that the body and environment do much of the cognitive work does not eliminate the need for internal states that function as representations; it simply changes their nature from passive mirrors of the world to active controllers of action. Thus, Clark stands as a pivotal figure who accepts the "hunger" for representations in complex cases while rejecting the "cognitive sandwich" model for simpler, coupled interactions.

2.3 The Pluralist Alternative

However, not all theorists sympathizing with dynamical systems approaches are so pugnacious. Some of them think that both formalisms can be useful and advocate for some form of methodological pluralism. Beer and Williams's (2014) study serves as a prime example of such an attempt at reconciling two different traditions. In the study, they have built a model agent that was asked to perform a relational categorization task. Importantly, it was an evolutionary model, so it lacked a predefined specification set by the programmers. Because of the diachronic nature of the agent's development (namely, its evolution), they applied the DST toolkit to describe it, considering it to be a trajectory with certain bifurcations, attractors, and limitations. The result is a 'state diagram' (sometimes enticingly called 'Beer diagrams') a map of temporal change of the system.

At the same time, they performed a classical information-theory analysis of the flow within the system. Using different 'mathematical lenses,' they 'have suggested that information theory and dynamical systems theory should be conceived as merely different mathematical lenses through which we can examine the operation of a cognitive agent. Each lens provides a unique explanatory perspective, distinctive language, and a complementary set of tools for framing and answering certain kinds of questions' (Beer & Williams, 2014: 23).

Importantly, though, this kind of reconciliation amounts to an extension of both IT and DST over their traditional scope and toolkit. For example, informational analysis usually concentrates on static relations (assumes relative stability of the system—in the developmental scale at least).[17] In the case of DST, there is a converse situation. Beer and Williams hence extend IT tools to measure not only traditional measures like mutual information during particular points in time but also to measure the flow over time and over individual stimulus values to track its gain, loss, and transfer across components.

Computational studies of this kind are bolstered by empirical work, too. To achieve that, Randy Beer cooperated with Izquierdo on klinotaxis (the process of successive comparison in stimulus intensity to achieve spatial orientation performed by very basic organisms) in the renowned *C. elegans* roundworm (nematode). Here the process of mapping information theory onto the dynamical landscape goes roughly as follows. Firstly, the connectome of the organism is mapped to represent (sic!) the pathways of potential information flow. Then the network of connectivity is established (functional based on the structural description). Finally, simulations are performed in order to see whether the prediction matches the behavior of the

[17] This is perhaps a controversial formulation. We can instead say that informational analysis usually abstracts away from temporal considerations.

target organism. According to Beer, Williams, and Izquierdo, IT and DST descriptions are complementary and mutually illuminating.

Scott Kelso (Kelso, 1995; Kelso & Engstrom, 2008; Rączaszek-Leonardi & Kelso, 2008) similarly advocates for this synthesis, formalizing the integration through the framework of Coordination Dynamics. He explicitly defines biological coordination as "informationally coupled self-organizing dynamical systems," thereby rejecting the traditional dichotomy that separates symbolic information processing from physical laws. For Kelso, the bridge between these domains is metastability—a dynamical regime where a system exhibits tendencies rather than fixed stable states, effectively creating a flexible landscape where information is continuously created, preserved, or annihilated. In this view, dynamics provides the necessary physical mechanism (the equations of motion), while information provides the functional context (the meaningful coupling), rendering them not merely compatible but, as Kelso and Engstrom posit in The Complementary Nature (2006), inextricable aspects of the same biological reality

This is in discord with more radical DST theorists such as Anthony Chemero, who claim that a complete dynamical description renders computational descriptions unnecessary. In their view, computational systems are merely a subset of dynamical systems, a particular case defined over discrete time and space, but are functionally irrelevant to biological intelligence. Chemero explicitly defines his position as "Radical Embodied Cognitive Science," arguing that "cognition is to be described in terms of agent-environment dynamics, and not in terms of computation and representation" (Chemero, 2009, 29). Zednik (2011) proposes two ways of reading these kinds of accounts, which he calls ‘thick’ and ‘thin’ interpretation. In the ‘thin’ version, the thesis of reducibility of representational (and a fortiori informational) states characterizes this relation in terms of describability: informational states can be redescribed in dynamical systems terminology. They are mutually translatable

without the loss of their original intentions. The thick interpretation assumes hard reducibility, meaning getting rid of tentative representational states entirely. Cognition is regimented by time synchronization. Dynamical states do not have stand-in relations; they are not in the business of relaying messages. They point out the fact that most sensorimotor processes do have dynamical descriptions and are perfectly explainable without appeal to representational language. But there is still considerably less evidence for the success of this strategy in representation-hungry problems, especially for making sense of how local internal processing happens. What, in my opinion, makes the hard anti-representationalism dubious, at least at this point. DST can be helpful, but not on its own.

An interesting side question emerges then: can DST be done this way without mechanisms? I do not wish to dwell too much on the mechanistic literature, but some comments are in order. Kaplan and Craver think that dynamical description is not explanatory, as it amounts just to 'mere description' (2011:602). Bechtel and Abrahamsen (2010) advocate for 'complementarianism'—the notion that dynamical description is concerned with global behavior of the system, not the local working of parts. As such, it needs to be buttressed by the mechanical description. It is controversial, though, how reconcilable these versions of dynamicism are with 4E accounts (which are direct continuations of DST in many senses, as we have seen). Especially the problem of extendedness is discussed. Can processes happening partly outside the 'scope' of a particular agent (with the use of artifacts, for example) still be considered a mechanism? Zednik is inclined to believe it is possible, but Bechtel disagrees, claiming this is a question of 'locus of control,' which directly forbids extending mechanisms this way.

To sum up, outside the trenches of 'representation wars,' there is a third camp of theorists who claim that representationalist and anti-representationalist accounts are not in direct conflict with each other. They claim that two traditions are either

mutually interchangeable or even complementary accounts of cognition, which work best for different timescales and scopes. Dynamic theory is well-suited to long-term temporal changes of the system, such as in the evolutionary models. It also works best on the scale of bigger, global networks. IT descriptions and representational accounts more broadly are better suited to shorter-scale and smaller problems, as they can explain why particular parts of the system are coupled together (namely because of their content), which remains a mystery for DST.

3. The Minimalist Defense: Deflationism

Instantly, a diligent objector might protest here. Why is deflationism included under the anti-representationalist auspice? There is some merit to that. Deflationism's goal is not to get rid of representations after all, just to make the term (or rather the object that is the term's target) more minimal. I hasten to clarify, then, that I discuss deflationism here as an opposition to a more rapacious representationalism, one that claims, e.g., that representations as labels are only pertinent to more sophisticated explanatory targets, such as one used by human perceptual psychology, but not in biology of basal cognition (e.g. Lyon, 2015) or robotics (Brooks, 1991). Deflationism is an attempt to salvage representationalism from anti-representationalist critique without going overboard (switching sides). This is a precarious position, as they have to prove that there is something more to cognizing than anti-representationalists propose, i.e., explanations in terms of dynamical systems or online couplings with rich environmental features. At the same time, they have to show that shaving down the inflated (in their understanding) classical representations gives them more explanatory purchase.

3.1 The Argument from Practice

There are at least three reasons that are put forward for this purpose. The first one can be called the argument from scientific practice. In neurobiology, researchers routinely say specialized neurons or neuronal ensembles that reliably fire together with a particular set of conditions are 'representations' of these conditions. Parts of the primary visual cortex that are selectively responsive to bars tilted at a particular angle are said to be representing this kind of position. Similarly, nodes in the neural network that are systematically correlated with the presence or absence of a particular feature are supposed to be 'representing' this feature within the architecture of the network[18].

Surely though, the mere fact the scientists speak and write in such matters is just not a solid argument on its own. It can be simply an infelicitous way of expressing or, more neutrally, simply a manner of speaking, kind of a jargon that is not very precise, as they do not care for this kind of conceptual precision in their nitty-gritty work. It can also be claimed that these are homonymous terms, but this would be too much, as there is a clear common denominator between practical and postulated usages. There is something important to be salvaged in the process of conceptual analysis, and some original intention of the term to be retained and made clear. Anyway, the argument from practice is far too little to do the trick.

---

[18] Of course, this is a gross simplification; sometimes the bigger ensemble of nodes can be used to represent a feature, and also the topography of the network (pattern of connectedness) and setting up of weights and parameters can play a role. But regardless of the minutiae of the postulated representing relation, it is assumed to act as a 'representation,' inasmuch as it is a reliable stand-in for some feature in question. Of course, even such a formulation is sometimes too rosy, as often (especially in the case of bigger, convoluted nets) there is no way to reliably delineate what represents what, which brings the famous problem of blackboxing and algorithmic obfuscation to the table (Mittelstadt et al.,2016)

3.2 Egan's Deflationism

A second strategy for defending deflationism is aimed at sidestepping the 'representation wars' altogether. This line of reasoning is championed by Frances Egan. This approach mixes a 'realist account of representational vehicles and a pragmatic account of representational content' (Egan, 2020:41). Contents have, according to this theory, only a heuristic value and do not serve an explanatory role. They are not specified by the representation relation but are mobilized for pragmatic reasons, as a sort of explanatory gloss. We can imagine the researcher saying that the particular part of the neuronal make-up of the organism is about a particular part of the environment, but only to make the statement more understandable for others in the community, for example. The content itself is not crucial for fixing the representation relation; vehicles are. Deflationism as such is the methodological stance aimed at eluding the charges against representationalism on the content side without giving up on their realist interpretation in the case of vehicles that are usually conceived to carry the content (Myin & Hutto, 2020).[19]

3.3 Facchin's Critique of Egan's Deflationism

However, this deflationary account suffers from two important problems, as identified by Marco Facchin (2022). First, by distinguishing between "cognitive contents" (the gloss) and "mathematical contents" (the real computation), Egan attempts to secure naturalism. But mathematical contents themselves face an indeterminacy problem. Facchin argues that mathematical contents do not satisfy important and widely accepted desiderata that all theories of content are called to satisfy, such as content determinacy and naturalism. If we cannot naturalize the

[19] Although, more precisely, Egan seems to reject the entire content/vehicle dichotomy, at least in the realist sense of the term, as is usually invoked in the literature on cognitive content.

mathematical content without appealing to the very semantic properties we tried to deflate—using cognitive content to fix which mathematical function is being computed—the account risks circularity. Secondly, Facchin claims that there are cases in which mathematical contents cannot play the explanatory role the deflationary account claims they play. He proposes an empirical counterexample to show that mathematical contents alone often fail to explain the system's failures. In cases such as adversarial examples in neural networks, the error is often only intelligible by reference to the cognitive content (e.g., classifying a panda as a gibbon due to imperceptible noise) rather than the pure mathematical function. Thus, deflationism may not be robust enough to carry the explanatory load.

4. Information (and/or as?) Representation

It was mentioned already that information-based theories are usually committed to representationalism. It is clear that representations are information-bearing states. They are contentful because they carry information about something. But we have not discussed in detail that kind of relation that links information to representation more precisely. This matter should be addressed.

There are at least two major strategies that can be employed. One postulates that information is representation *simpliciter*. They may not be all the representations there are (and almost certainly are not), but they are representations anyway, plausibly in a minimal sense. This strategy seems like a difficult one, as Shannonian theory is famously stringent. But recently, the challenge was accepted by Martinez (2019a), who claims that Shannonian theory is a formal theory of representation. This is a minority but nevertheless soundly motivated position. The second way to adjudicate information-representation entanglement seen throughout the

literature much more often is to say that representation is information plus some other factor (or maybe multiple factors). Per analogiam with the well-known situation in the post-Gettierian epistemology of knowledge where theories that build on the traditional 'justified true belief theory with additional conditions are called 'JTB+X' theories, we can call this strategy ***I+X*** *theories.*

4.1 Martinez's Theory

Let's move to the first strategy first and inspect Manolo Martinez's proposal more in-depth. Martinez's mode of operation is simple. He proposes to take a set of criteria laid out for representations and then try to argue that processes already present in Shannon's theory of information already match them. To do so, he summons Birch and Smortchkova's conditions for representation. There are three of them.

(a) correlation between the detected feature and the pattern of neural activity is exploited in further processing;

(b) this further processing ends up guiding action and explaining success; and

(c) misrepresentation is possible (e.g., when neurons fire in the absence of the stimulus they evolved to detect), and this helps explain unsuccessful action. (Birch & Smortchkova, 2019:19).

Representations are then not only correlations, but exploited correlations that allow the agent to perform effective action and also provide room for misrepresentation when there is no success. Martinez argues that source coding and channel coding suffice to clear all three bars. The first one is the easiest, as this requirement is actually assumed before utilizing the formalism. In a sense, the assumption that the signal is used further on (i.e. processed downstream) is what

allows for source-channel analysis to be performed. The subsequent usage by some decoder is a prerequisite of considering it to be an encoded signal.

The second issue is tackled through the notion of reliability of transmission. Martinez claims that the distortion rates as defined and measured by IT are tools aimed at assessing the rate of faithfulness. We can interpret this rate of faithfulness as a measurement for successful communication. In the case of an artificial system, this is set manually by the constructors and programmers using predefined benchmarks of satisfactory performance. But for natural systems, distortion measures constrain how the target information processing system helps increase the fitness of participants in the communication (Martinez, 2019a:12).

The answer to the third challenge follows naturally from that interpretation. Misrepresentation is then the very distortion of the transmission. Quantifying the faithfulness of the transmission is, in this sense, the measurement of both representation (how well the original message was retained) and misrepresentation (how much information was lost along the way). While one might argue that distortion is not necessarily misrepresentation—since a distorted copy is still a copy, whereas misrepresentation (like seeing a zebra as a horse) involves error rather than noise—Martinez's account provides a powerful deflationary toolkit.

I am, in fact, highly sympathetic to Martinez's remarks of underlining the role of compression-like processing in representational schemas. The evidence he mobilizes, coming from psychoacoustics (perceptual coding—Brandenburg, 1999), linguistics (learnability—Kirby et al.,2015)[20], and machine learning (dimensionality

[20] The similar ideas were investigated experimentally by Steinert-Threlkeld & Szymanik (2019). They claimed that particular linguistic categories that are more easily reducible mathematically (defined as 'convex') are demonstrably easier to learn. For that reason, they presented evidence both from machine learning and human learning, where convex categories were learned much faster than non-convex (intuitively: heterogeneous).

reduction—Tenenbaum, Silva & Langford, 2000), is compelling. The discussion of efficient coding in the brain and principles of cognitive economy (on higher scales) is, in my opinion, strong evidence for the importance of compressive mechanisms in representational features. I build my own version of this position in chapter 3 and review different bulk empirical evidence for that. Martinez's emphasis on the integration of independent sources of information further strengthens this view, suggesting that what we call "content" may often be the result of optimal source coding strategies designed to maximize fidelity under metabolic constraints.

However, despite the elegance of Martinez's reduction, most theorists agree that we need more conditions that representations have to have. There are several important properties that are thought to be essential or at least important for representing.

4.2 Skyrms and the Evolution of Vectorial Content

The traditional philosophy of mind has usually been dominated by the propositional view of content. On this view, a representation of any kind, whether a sentence, a belief, or a neural state, is a vessel that carries a specific proposition p (e.g., "There is a leopard"). This proposition has truth conditions: if the leopard is present, the representation is true; if not, it is false. We all know the story. While this binary logic works well for human language, Brian Skyrms (2010) argues in *Signals: Evolution, Learning, and Information* that it is a poor fit for the biological reality of signaling systems. Hence, he does not concern himself with conversing with the tremendous baggage of said philosophy of mind. He explicitly claims that most of the current philosophical approaches have been unproductive and suggests just brushing them aside.

As a result, Skyrms proposes a naturalistic alternative based on evolutionary game theory. He demonstrates that we do not need to presuppose "intentions" or

"concepts" to explain the emergence of meaning. Instead, meaning evolves spontaneously from the dynamics of interaction. Using the framework of Lewisian signaling games, Skyrms models a population of agents: senders (who observe a state of the world) and receivers (who perform an action). Through simple reinforcement learning (e.g., Roth-Erev dynamics[21]) or replicator dynamics, these agents spontaneously converge on stable signaling equilibria and avoid getting trapped in the ruts and valleys of sub-optimal pooling with no way of adapting further.

That's certainly commendable. But what could vectorial content possibly mean? Crucially, Skyrms argues that the content emerging from these dynamics is not propositional but *vectorial*. In a signaling game, the receiver starts with a prior probability distribution over the states of the world (based on the base rates of events). When the sender emits a signal, the receiver updates this distribution to a posterior probability. Skyrms defines the informational content of the signal as the vector of this probability shift.

Formally, if the state space consists of three possibilities {S1, S2, S3}, the receiver's prior might be the vector ⟨0.33, 0.33, 0.33⟩. Upon receiving signal M, the distribution shifts to ⟨0.9, 0.1, 0.0⟩. The content of M is not the sentence "It is S1"; rather, the content is the vector itself (or the mapping from the prior to the posterior).

This shift from propositions to vectors is not merely a mathematical convenience; it allows Skyrms to solve the problem of "half truths" and ambiguity. In biological

---

[21] Roth-Erev learning is a reinforcement learning rule where agents assign weights to actions proportionally to accumulated past rewards, then select actions probabilistically based on these weights. The key innovation is the discount factor $\lambda \in [0,1]$, which implements "forgetting"—gradually reducing the influence of older experiences on current decisions. This enables agents to remain adaptively responsive (ergodic) rather than becoming locked into suboptimal behaviors, as occurs with basic cumulative reinforcement ($\lambda=1$) or requiring exponentially long convergence times with pure randomization ($\lambda=0$). In signaling game simulations, Roth-Erev with forgetting (λ between 0 and 1) achieves 100% convergence to optimal communication conventions, where traditional approaches often fail (Catteeuw et al., 2011; Roth & Erev, 1995).

environments, signals are rarely perfect indicators. A vervet monkey alarm call might indicate a high probability of a leopard, a low probability of a snake, and zero probability of an eagle. A propositional theory struggles to classify this: is the signal "Leopard"? "Snake"? "Predator"? It forces a binary choice where none exists. Skyrms' vectorial view captures the reality: the signal means exactly the probability distribution it induces ($\langle 0.8, 0.2, 0.0 \rangle$). The signal is an accurate representation of the *probabilistic structure* of the environment, even if it does not guarantee a single outcome.

Skyrms also addresses how such systems can expand[22]. He identifies information bottlenecks as the drivers of complexity. When the number of relevant states in the world exceeds the number of available signals (or the receiver's ability to discriminate), the system reaches a plateau in utility. Skyrms models the "invention" of new signals via the Hoppe Urn model (or the Chinese Restaurant Process)[23]. By introducing a "mutator" (a small probability of inventing a new signal), the system can break the bottleneck, partitioning the state space into finer grains. While I will argue in Chapter 8 that this stochastic mechanism is insufficient for *online cognitive generativity*, it provides a robust account of the *phylogenetic* origin of the code. It explains how the primitive vectors of the neural code are initially carved out of the noise.

---

[22] Which is interesting for later considerations about generativity: how do new informational states emerge?

[23] It's an intuitive way to visualize how new signals are invented. Imagine a restaurant with an infinite number of tables where customers (signals) arrive one by one. Most customers choose to sit at an occupied table with a probability proportional to how many people are already there, representing the reinforcement of existing signals. However, occasionally a customer chooses to sit at an empty table instead, representing the invention of a brand-new signal. This ensures that while popular signals dominate, there is always a non-zero probability of creating a new one to solve a new problem

Despite its mathematical elegance, Skyrms' account faces significant criticism for being *too* permissive (as mathematically elegant theories tend to). Kim Sterelny's review, "A Glass Half-Full: Brian Skyrms's Signals", provides a critical examination of Skyrms's formal models of signaling. Sterelny finds much to praise in Skyrms's simple and elegant approach, which demonstrates that signaling can emerge spontaneously in simple Lewis signaling games, even in the absence of elaborate cognitive capacity or pre-existing communicative abilities. However, Sterelny ultimately concludes that the models' simplicity pushes them and the real-world exemplars apart. His central philosophical caution is that Skyrms, by prioritizing an internal model-building strategy that abstracts away from biological constraints, may limit the models' ability to explain the actual natural history and constraints on coordination and communication, such as the constraints on specialization or the problem of how a reward is coded in a subcognitive learning system during escape from danger.

Furthermore, Sterelny argues that the model's assumption of a forced choice between distinct, state-specific actions loads the dice in favor of the emergence of signals. Specifically, he introduces the possibility of the "hedgehog response," a viable, general-purpose strategy (e.g., a hedgehog rolling into a ball) that pays reasonably well across all possible environmental states without being optimal in any. If such a strategy is available, an agent only has an incentive to learn to tune its behavior to specific world states through signals if learning is quick and cheap, or if the payoff difference between the optimal strategy and the hedgehog strategy is large. Therefore, the availability of a decent, all-purpose response may block the spontaneous emergence of signaling, revealing a limitation of the Lewis game as a base case.

There are other worries, too. In the review of *Signals* Peter Godfrey-Smith (2011), expresses doubts over Skyrms's potentially overly permissive definition of

information. He argues that Skyrms conflates natural signs with intentional signals. Under Skyrms' definition, anything that shifts a probability vector counts as a signal. We can clearly see why it's problematic. A dark cloud shifts the probability of rain; therefore, the cloud "signals" rain. While true in a broad informational sense, Godfrey-Smith argues this fails to distinguish between mere physical correlation and functional representation. In cognitive science, we need to explain systems that represent the world *for the purpose* of guiding action, not just things that happen to correlate.

Furthermore, Millikan points out that Skyrms' vectors lack a distinct *direction of fit*. High-level cognitive states are usually classified as either indicative (describing the world: "The door is open") or imperative (directing action: "Close the door"). Skyrms' primitive signals are "pushmi-pullyus" (Millikan, 1995)—they simultaneously map the state space and trigger the action space. The vector ⟨0.9,0.1⟩ tells the receiver "It's probably state 1" AND "Do action 1." While this works for bacteria or reflex arcs, it fails to capture the decoupled nature of higher cognition, where information is stored (descriptive) without triggering immediate output (imperative).

The Vectorial Thesis seems like a natural ally for trying to marry informational content with neural representations. If the contents are fundamentally probability vectors, not sentences, this is interpretable in terms of dimensionality spaces that neuroscience routinely operates on. This also aligns with our putting up the format requirements in their rightful spot (as is to be extensively argued for in Chapter 3). However, I contend with Millikan's critique: the vector alone is not enough. To verify that a vector is a *representation* and not just a cloud-like correlation, we must apply some sort of teleosemantic constraint. This is why I think it is crucial to juxtapose him with Martinez's ideas. The vector must function as a vehicle within a specific consumer architecture. Skyrms gives us the geometry of the content; the consumer gives us the status of the representation.

However, the true utility of Skyrms' framework for the Neo-IPP becomes evident when defending neural representations against the eliminativist charge. In a prominent critique, Brette (2019) argues that the very concept of a "neural code" is a fallacy because neurons lack access to the external world—they are trapped in the skull and cannot verify the "truth" of their signals. Without an external homunculus to read the code, Brette claims, neural firing is merely causal, not representational.

However, Cao and Rathkopf (2019) leverage Skyrms' evolutionary logic to dismantle this objection. They concede that Brette successfully refutes "immodest codes"—symbols that require a predefined lookup table or external verification. But they argue that the brain relies on "modest codes": signals that acquire meaning solely through their functional role in coordinating internal activity. Invoking Skyrms (2010), they demonstrate that "action policies can endow bare causal commerce with meaning". In Skyrms' models, agents evolve meaningful signals (e.g., distinguishing "predator" from "food") purely through trial-and-error interaction and differential reinforcement, without ever conceptually "knowing" the external facts.

Applied to the brain, this implies that neurons do not need to "see" the outside world to represent it. Like Skyrms' agents, they develop stable firing policies that coordinate downstream action. As Cao and Rathkopf vividly picture, "[Skyrms] have developed mathematical models showing how action policies can endow bare causal commerce with meaning." Over time, these signals acquire semantic content because they are selected for their success in guiding the organism's behavior. This is why the BTBI rats needed a structured regime of fixing signals to behaviour - 'acquiring semantic content.' Thus, by adopting the Skyrmsian view, we can define a neural code that is intrinsic to the system—generated by its own internal signaling games—rather than one projected by the experimenter. This secures the

status of neural representations as real, evolved structures that are causally activated in constant tokening and transitions, immune to the eliminativist critique.

4.3 Hiding in Plain Sight. Alistair Isaac on 'Semantics Latent in Shannon Information

While Skyrms arrives at the vectorial view through the biology of evolution, Alistair Isaac (2018) arrives at a strikingly similar conclusion through the formalism of information theory. In his pioneering paper *"Semantics Latent in Shannon Information,"* Isaac attacks the pervasive dogma in philosophy of mind that Shannon information is purely syntactic—a measure of quantity (bits) devoid of meaning.

Since Shannon (1948) explicitly stated that "meaning is irrelevant to the engineering problem," philosophers like Dretske (1981) and Floridi (2011) have assumed that information theory can quantify *how much* data is sent, but not *what* it says. On this view, a 10MB file of random noise and a 10MB file of Shakespeare contain the same amount of "information" (entropy), proving that the theory is blind to content. This causes a sort of semantic void to open up. Philosophers have struggled mightily to explain that. Isaac was probably the first to tackle this head-on (after 70 years!).

And he does it in an unexpected fashion, but in a somewhat similar move to Martinez and Mann. They both claim that there is simply more in classical IT than it purports to have. In a way, they disregard the warning of Shannon himself when he admonished those who were to overextend his formalism. For long, to yield to this temptation was to heed the siren's signal.

Isaac overturns this by arguing that semantic content is mathematically latent in the measure of Kullback-Leibler (KL) Divergence (also known as relative entropy). He introduces the concept of "S-vector semantics." Which can be understood as follows:

When a sensor R reads a signal r, it does not just reduce uncertainty; it transforms the probability distribution of the source S. From:

Prior: P(S) (the distribution before the signal). To:

Posterior: P(S|r) (the distribution after the signal).

Isaac defines the content of r not as the scalar value of the uncertainty reduction (bits), but as the vector of the update itself.

$$v = \log\left(\frac{P(S)}{P(S \mid r)}\right)$$

This vector encodes exactly how the probability mass has shifted across the entire state space. If the signal is "smoke," the vector shifts probability mass drastically toward the state "free" and away from "clear air." Crucially, this vector captures the structural relationship between the signal and the world.

The neat feature of the S-Vector theory lies in its elegant solution to the problem of the "Disjunction Problem" (and thus to misrepresentation challenges). This is the classic objection to informational theories: If "smoke" is caused by "fire," does it mean "fire"? What if "dry ice" also causes "smoke"? If the signal occurs in both cases, does it mean "fire OR dry ice"? If so, the signal is never false, just vague. We lose the ability to say the system was "in error."

Isaac's S-Vector solves this by decoupling information from truth. Consider the "Dry Ice" case. The system detects smoke. Based on the channel statistics (where fire is the cause 99% of the time), the S-vector shifts the probability of "fire" to 0.99.

The content states, "Fire is highly probable" (the vector ⟨0.99,0.01⟩), whereas in reality, dry ice (the state ⟨0,1⟩).

The signal carried accurate information. Given the statistical laws of the channel, the update was correct: smoke *does* raise the probability of fire. The fact that the single-instance outcome was dry ice does not invalidate the informational content. The system was "misled by the truth"—it received valid probabilistic information that happened to diverge from the specific instance. This allows us to define error not as a failure of the signal to exist, but as a divergence between the informational vector (the representation) and the actual state.

4.4 Rung on a Ladder

We can now move to discussing the second type of strategy—the I+X version. Mark Sprevak in a recent article (2020) writes, 'What is the relationship between information and representation? Dating back at least to Dretske (1981), an influential answer has been that information is a rung on a ladder that gets one to representation. Representation is information, or representation is information plus some other ingredient.' (Sprevak, 2020:591).

Indeed, it was Dretske who pioneered the attempts to explain representation using the I+X strategy. He acknowledged the centrality of this relation for any serious informational epistemology, so he starts his Knowledge and Flow of Information with the definition of what he calls 'natural representation' right off the bat. "The fundamental idea is that a system, S, represents a property, F, if and only if S has the function of indicating (providing information about) the F of a certain domain of objects" (Dretske, 1981:2). We can see it is quite a simple definition. It consists only of indication and functional components. The latter is the sole 'additional' property. As such, Dretske's notion applies to both mental and non-mental representation.

For example, a speedometer is a representational system on his terms, as it is an indicative system with a function of providing information about the speed of the vehicle. The indication happens when the system can change its states in a manner

systematically corresponding to the changes of the definite values of the indicated property. The position of the needle relative to 0 corresponds reliably to particular values of speed. Thus, a speedometer is a natural representation, regardless of the potentially misleading qualifier.

There is a second potentially unintuitive consequence of Dretske's account, coming from another side, in a sense. It is, under scrutiny, more mentalistic than the 'natural' label might suggest.[24] Since the system has to have a particular function of indicating some property to be representing it (rather than merely indicating), some features that it also might inform about contingently are not represented. Let me give an example. Our speedometer can also be said to quite reliably inform us about the state of the gas tank. The more we speed, the higher the fuel consumption. A more inclined needle indicates a sooner visit at the petrol station. But the speedometer's function is not to monitor the level of gas.

The concept of Dretskean representation is intentional and, as such, more mentalistic than we would perhaps expect from the openly naturalistic account (McIntyre, 1999:2). Of course that does not mean that the 'proper' natural meaning of representation is excluded. Natural systems just achieve their function due to the challenges encountered 'out there' in the world that they are aimed at solving (or helping to solve). The better they help with these challenges, the better they are performing their function and improving the overall 'fit' to the task. This pressure is generated by evolutionary and developmental processes. 'Face cells' in the fusiform face area represent—well, faces because they reliably indicate their presence (and structure) according to their function. This information is passed on to the further chains of processing (downstream).

---

[24] Dretske also acknowledges the notion of 'mental representation' as opposed to 'natural,' so the mentalistic connotation of the latter is perhaps problematic.

Natural mental representations are then, in Dretske's sense, in the head. But that should not make us jump to conclusions that his account is internalist. On the contrary, he openly claims, 'The Representational Thesis is an externalist theory of the mind. It identifies mental facts with representational facts, and though representations are in the head, the facts that make them representations—and, therefore, the facts that make them mental—are outside the head. A state of the brain represents the world in a certain way only if it has an appropriate information-carrying function. Since functions have to do with the history of the states and systems having these functions, mental facts do not supervene on what is in the head. What is in heads A and B could be physically indistinguishable, and yet, because these pieces of gray matter have had relevantly different histories, one is a representational system, the other is not; one is the seat of thought and experience, the other is not.' (Dretske, 1981:124-125).

Representational relation is not fixed purely by the internal states of the representing system. There can be nothing intrinsically declarative for establishing what is being indicated but not represented. Indicating relations could have this feature as internal relations of system states; for example, ratios of angles on a speedometer's dashboard are in strict correspondence to the incremental changes of the target property: velocity. But representation is begotten by function, and it cannot emerge from within[25]; it needs external scaffolding, be it convention of design (in artificial systems) or developmental/evolutionary history. The consumers of representation, while in the same informational pipeline, are outside recipients. It is the external environment, too, that eventually provides verification of how well the function is performed.

---

[25] Inside the particular system, not necessarily inside the skull, for instance. For the fusiform gyrus (located in the inferior temporal cortex), the thalamus, the recipient of some information, is 'outside.'

To summarize, Dretske's account for representation is information transfer, but only defined functionally and bolstered with an externalist frame. It is quite minimal in the sense that it does not postulate many additional properties to mere information processing but is still more restrictive than Martinez's account that tries to make do with just an informational toolbox. Nevertheless, according to many theorists, the Dretskean proposition is still too narrow, and we need to account for more properties of representational states.

5. Constitutive Properties (The "X" Factors)[26]

In this section I review the various claims about the 'additional' properties of information scattered around the philosophical literature. They are additional in the sense of openly going further than minimalistic or engineering-like accounts that (if expressed in explicit terms at all) simply assume that any standing in relation is sufficient for representing, as we have seen before. They are in this sense more restrictive than postulates of mere information carriage or statistical and lawful covariance as sufficient for representational status. Thankfully, there is a plethora of these kinds of supplements on offer.

First, I discuss the potential merits of these additional properties and the motivation of their employment one by one. Then, after such a tour, I conclude with a discussion of whether some, all, or none of them can be considered as necessary

---

[26] It is worth noting that the criteria for representation proposed by Birch and Smortchkova (2019)—exploited correlation, action guidance, and the possibility of misrepresentation—align partially but not fully with the 'X-Factors' defended here. Their 'minimal' account effectively captures the functional profile of what we term 'Online/Coupled' states (see Section 5.3), accepting action-guiding correlations as sufficient for representation. However, our framework imposes the stricter structural requirements of Decouplability (5.2) and Mirroring (5.4) to distinguish genuine, portable representations capable of offline processing from mere specificational states or sensory-motor couplings

features of all possible representations, or maybe whether their scope is limited and more specialized.

5.1 Non-Ubiquity

I want to start with perhaps the most trivial of all of them—the flight from omnipresence. Representing is in many ways truly awe-inspiring. It is at the most intuitive level at least successful extraction of information about the environment, but also mirroring it somehow with distortions as minimal as to successfully further one's goals. That's no small feat. This is why there is an important meta-theoretic conviction among many theorists that what counts after our theory of representation should not come as too common. That is in stark contrast to information, which is supposed to be very common—both in quantity and accessibility. There is ample information in the world (even if one is not an ecological psychology proponent), and all the agents, both living and artificial, are uncontroversially nifty processors of it. This is perhaps put quite loosely, but the same thing was expressed more succinctly by Orlandi (2015:105), who wrote that 'while information is ubiquitous, representation is not.

There is also a deeper worry about the overly promiscuous account of representation than this intuitive understanding of its exclusivity. It is a methodological argument against concepts inflated to the point of triviality. We have already seen the thrust of this argument in Burge's comments against accounts of information as statistical co-variance. The argument is roughly as follows: by overextending the concept's boundaries, we discard what afforded it with particular explanatory interest in purchase. In some sense, there is nothing wrong with this kind of definition, but then we need some new term to keep the relevant distinction in place. As long as some line of division (demarcation) exists, it is fine. But for pragmatic and communicative reasons, it should be avoided. But for

pragmatic and communicative reasons, it should be avoided. After all, the permissive policy about the term buys us nothing explanatorily, just moves the problem one step further (to the new distinction) and confuses both the existing participants (as the call for liberal usage usually effects the shift in its semantic field) in the debate, but even more so the newcomers.

'Inference' serves as a useful example of such a term. Originally, it had a strongly mentalistic, high-level, propositional interpretation as a personal-level procedure usually afforded by language-like representations governed by logical relations. However, firstly, the unconscious inference movement in perceptual psychology, was then followed by the New Look psychology, which extended the scope of inferential processing to early-level perceptual acts. Then the emergence of Bayesian and predictive accounts of cognition further obfuscated the view, pushing the understanding of inferential processing as a very general mathematical abstraction. In some ways this process was understandable and even laudable. For philosophical work on inference was marred by hyperintellectualization[27], which effectively excluded any possibility that empirical progress could impact the debates. Work by Friston and colleagues pushed the boundary even further, ascribing the notion to even the most primitive agents (Friston, 2013; Pezzulo et al., 2022). The willingness to 'unificate' cognition on all levels using an allegedly extremely applicable mechanism (in fact, - formalism - Miłkowski & Litwin, 2022) rendered the original notion trivial. To restore original debates in their actual axes of disagreement, some new labels emerged for the high-level processing governed by logical schemas - e.g., reasoning (McHugh and Way, 2018).

---

[27] I do not speculate here on the relation between the two and whether they should be kept separate or not, as it is not essential for illustrative purposes here (for discussion, see Rescorla, 2015).

To make the matter of potential conceptual confusion worse, there is ample usage of ‘cognitive’ and ‘perceptual’ representation in the literature pointing at the fact that there are important philosophical distinctions to be made between perception and cognition. Setting up the divide is an ancient philosophical problem dating back at least to the problem of empiricists' struggles with justification of perceptual states (but probably even to Aristotle’s theory of mind - see *De Anima* III.3, 427b); Everson, 1997). Recently, interesting new ideas emerged, for example, Beck’s proposal of the ‘stimulus dependence criterion’ (2018). The question is also frequently discussed by Ned Block in his relentless, platonic pursuit of identifying the ‘joint in nature’ between perception and cognition (2023).

What I will be interested in here are conditions for representing in general. However, it is clear that the most interesting application for my considerations is cognitive representation, as proved by the proposition to move the debates to the current state of neuroscientific knowledge. So when I will be talking about representation, I will mean cognitive representation in a broad sense—encompassing both potentially different perceptual and cognitive representation. If it happens to appeal to the more precise distinction (e.g., cognitive-perceptual divide) and hence cognitive representation in a narrower sense, I will underline this fact clearly in the respective passage.

This will, for example, be the case in discussing Nico Orlandi’s (2014) sensory representations and rule-based representations. This distinction is not merely terminological but structural: sensory representations may track "wired" regularities in the environment (wired transitions), while rule-based representations track "represented" regularities that allow for decoupling and inference (representational transitions). It bears directly on the *Taking Condition* (Boghossian, 2014)—the idea that a system must "take" a rule as a guide for action, rather than merely embodying a regularity—an important point that will come back

to help us defend commitments of IPP both in Chapter 5 and also in the last chapter on operationalizing information processing, namely in defining generating information.

5.2 Decouplability/detachability

When we were discussing how an informational relation emerges between cognitive structures and its targets, we said that there is a 'coupling' of, say, sounds of particular pitch and cells in the auditory cortex fine-tuned for its registration. The proponents of the interpretation of C-states as representations would plausibly say that the tokening of this representation was triggered by the pitch in question. At the same time, they would concede that this activation would be sustained only as long as the stimulus happens, is accessible, and is received by the agent. In this sense, the coupling we are talking about is strict—the activation is happening only when the pitch occurs, lasts (roughly, there is probably some delay for processing and cooldown) as long as it is maintained, and, importantly, it does not happen when its direct causal antecedent does not obtain.

This is, however, in stark contrast with the nature of our paradigmatic examples of representations. I can think of a shoebill stork even though I have (sadly) never encountered one in my life. I can imagine myself going to South Sudan and seeing one in the future, or even planning to do so. The same applies for counterfactuals. Shoebill storks playing football for Legia Warsaw? No problem whatsoever. There are a host of significant and common cognitive acts that are instantiated without being directly caused by the environmental features they are supposed to be about. They are decoupled from its purely statistical covariants.

This idea is not a new one. James, when discussing the differences between intentional and non-intentional behavior, summons the case of iron filings being attracted to the source of magnetic pull. When the barrier is established between

the filings and the magnet (we put a plank), there is no attempt on the part of the filings to circumvent the barrier. This is entirely different from Romeo's striving to reach Juliet, which cannot be stopped while separated but for a minor wall. He can scheme and push with his intentions even when there is no Juliet in the vicinity (James 1890/1950:5[28]).

Surely, there can be many in-between cases, and they are discussed in the literature, too. Sticking to the ferromagnetic world, there are known cases of magnetosome bacteria, which are pulled to magnets just as the filings are, while being living systems. They still are very poor at overcoming barriers, so they are probably not intentional systems, though. Moths following light are more skilled at circumventing obstacles, so they would constitute a more refined, more in-between case. And so forth, we can probably build some kind of spectrum between the border cases. But even the existence of a spectrum does not defeat the purpose of setting up a line of division (even if it blurs it when approaching it). Clark and Grush call this type of boundary 'fuzzy but important' (1999:8), and I agree. This is why the idea of decouplability as a hallmark of representation has still been raised by many in philosophical literature in cognitive science (e.g., Chemero, 2009, Clark, 1997, Sterelny, 2003). Gärdenfors (2003) calls representations that can be tokened without the target present (or cued) 'detached' representations.

So how can we advance it and make it better understood? An influential account was offered by Brian Cantwell Smith (1997). Cantwell Smith's goal was (in crude terms) to offer a theory of the evolution of simple communicative systems in such a way as to arrive at naturalistic semantics. He was then interested in how simple 'tracking states' can somehow get upgraded into representations somewhere along the way of cognitive systems getting more sophisticated evolutionarily. Tracking

[28] In memorable Jamesian parlance it is 'the pursuance of future ends and the choice of means for their attainment'.

states are coupled states and, as we saw, are cheap. Magnetosomes have tracking states. But more sophisticated organisms rely on them all the time, too. Especially when they can do the job just this way - it is a fairly reliable process after all. An observing hawk, an echolocating bat, or a sniffing hound are all relying on tracking states.

But Cantwell Smith is interested in complexifying this picture, so he invites us to think about the rabbit, which keeps an eye on a prowling coyote. Rabbit is (potentially[29]) being hunted, so it tracks the predator closely. Then imagine that the coyote goes beyond the rock, breaking the line of sight between the hunter and the prey. The distal stimulus of the coyote's ragged physique no longer impinges on the rabbit's retina via transmission of light (now occluded). But the rabbit still is extremely interested in the position of the coyote. Coyote is now beyond 'effective reach' (Cantwell Smith's lingo), but the rabbit is not helpless by any means. It can quite reliably guess where the predator may reappear and when. It can succeed even when the breaking of sight lasts for longer or is complicated by another contingency (it tilts the head to scan the other side of his surroundings, for example). The original object is not present, but also, in a sense, removed even farther. The spatial connotation is not accidental. Cantwell Smith proposes to call this property *distance*. He claims that true representations are born somewhere along this distance from 'effective discriminability' of tracking states.

A very similar account is offered by Sterelny in his *Thought in a Hostile World* (2003). He also sketches an evolutionary story about the move from mere detection to

---

[29] This is, incidentally, an important problem. Whether the decouplability of representation has to be actually decoupled or can be just potentially decouplable. Clark (1997) contends the latter suffices, but Orlandi (2015) thinks that actualization is needed.

'robust tracking'[30] (in his terminology, the counterpart to Cantwell Smith's 'effective coupling') and finally to *decoupled* representation. And similarly to the distance condition, he considers breaking with the presence of sensorily discriminable features of the environment to be crucial. The more removed the processing is from this input, the more sophisticated representation it (at least in principle) requires. Representation is mobilized (and accordingly should be in our theories) when no simpler system is needed. After all, 'the whole point of having a representation is that it can be used even when the representational object is not "reliably present and manifest" to its user' (Haugeland, 1998).

This line of thinking was continued by Orlandi (2014), who also argues forcefully for decouplability as a major distinctive feature of representations. They put great focus in their discussion on decouplability as 'coordination with absence.' They think representational systems are representational by virtue of being able to still coordinate their behavior (and modulate action guidance) even when the causal loop with the direct environment is broken. This explains why there appears to be a common denominator for some of the leading representationally hungry problems in cognition. Counterfactuals, inferences, imagination, recall, and future planning are all cases of tasks where organisms have to work with information that targets are not presently causally active - in other words, they coordinate with absence.

But absence is not a universal feature of representation-inducing situations. We plausibly form representations also about objects that are presently impinging on

---

[30] Interestingly, 'robust tracking' appears very early phylogenetically, according to him. Even bacteria are afforded this power, along with many more evolutionarily advanced phyla, including plants, arthropods, and vertebrates. For a more detailed discussion of the minutiae of Sterelny's theory and relations between detecting, robust tracking, and decoupled representations (see: Christensen, 2010 a critique). Especially the diagrammatic representations of relations between three core concepts are illuminative for understanding potential relations between them (e.g., how decoupled representations can emerge from both detecting and robust tracking and how the causal history is different in these two distinct cases).

the sensory apparatus, just not about all their properties. The objects in our visual field, for example, are often seen in constantly changing conditions. We move, tilt the head, the light sources change position of intensity, the objects themselves move, etc. But within this chaotic stream of fast-changing conditions, we tend to observe stable properties of objects. For instance, the shape of perceived objects is relatively stable regardless of the subtleties of depth cues. Of course, there is a breaking point when the perceptions change; the perceptions are subject to update (or even verification, depending on one's stance on the justificatory status of perception), after all. But the main point stands - we observe stable percepts, not a dance of ever-so-slightly-changing objects.

Orlandi argues that this phenomenon is explainable by the traditional (in perceptual science) distinction between proximal and distal features. Proximal features are the ones picked up by the perceptual system in microscale and susceptible to rapid changes - light intensity is a quintessential example of proximal property. Distal features, on the other hand, are the ones that are stable and quite impervious to changing conditions. Being red or being a dodecahedron are these kinds of properties. How can this stability happen, though? Orlandi thinks that the organism 'takes' the objects to be stable whenever it can employ representational processing. How else to explain this?

Fodor (1984; ch. 4) tried to use the criterion of nomicness. He argued that the distinction between properties picked up representationally and non-representationally should fall along the lines of whether the property in question is governed by natural laws or not. The motivation for this move is similar to Orlandi and Burge's - to avoid a reductionistic failure of panrepresentationalism. But the choice of criterion is a suspect one. Fodor thinks that all the properties for which registration can be explained purely by the nomic relations ought to be excluded. That is the thrust of his famous attack against thermostats and

paramecia. The behavior of both of them is accountable for by the laws of thermodynamics and mechanics.

However, the nomic account runs into trouble rapidly. After all, the existence of laws in science is currently hotly objected to (e.g., Cartwright, 1983[31]). But even if one believes in them, special sciences, such as biology[32] or even chemistry, are usually not deemed to be law-governed[33]. That would deem them falling under the representational domain, which would in turn render the entire account wide-representational, in direct contradiction of Fodor's original purpose and assumption.

Distalness is geared toward the same direction - the exclusion of properties that can be accounted for more minimally and preserving the explanatory integrity of 'true' representations. It succeeds in avoiding the pitfalls of nomic accounts, which

---

[31] I refer here to her important book *How the Laws of Physics Lie*. Cartwright's later work (with Pemberton and Wieten; 2018) presents a more nuanced position than the simple "laws lie" assessment suggests. While her earlier work emphasized that fundamental laws are true only of idealized models and not the messy real world, this paper argues that laws play a crucial explanatory role within mechanistic explanations of ceteris paribus regularities. Specifically, the regular behavior produced by a mechanism is what it takes for general principles governing the mechanism's parts to be instantiated together in their particular arrangement. Thus, mechanistic explanations are actually a subset of covering-law explanations, and their explanatory force derives partly from showing that the behavior is what must occur if the relevant laws are not violated. This suggests Cartwright sees laws as essential to understanding why mechanisms produce their characteristic behaviors, rather than dismissing their explanatory role entirely.

[32] Schaffner argues that biology does have laws, though these are "provisional generalizations" that are narrower in scope and subject to more exceptions than physical laws. For Schaffner, biological explanations often involve "overlapping interlevel theories" connecting phenomena at different organizational levels and are fundamentally mechanistic—explaining through causal mechanisms rather than strict deductive subsumption under universal laws. This makes biology more aligned with a "models and mechanisms" approach than a covering-law model.

[33] The explanatory autonomy of chemical laws is contested: Robin Hendry (2010) argues for their distinct status via "strong emergence," positing that molecular structure exerts downward causation on sub-molecular processes. Conversely, Hinne Hettema (2012) maintains that chemical laws remain ontologically reducible to physics, viewing the lack of strict derivation as a mathematical approximation rather than evidence of independent chemical causality.

are too narrow, as we have just discussed. They are also subject to renegotiation if the empirical evidence changes, though. This is surely a positive, but it is important to bear in mind that which properties are distal is ultimately an empirical question. The way to ascertain it is roughly to find which perceptual properties are stable in the environment and can be taken advantage of without the appeal to constructive processing (such as in traditional explanations to the 'inverse problem' or the 'underdetermination of the stimuli'). Natural scene statistics (Oliva & Torralba, 2001) is one way of exercising this strategy in practice. According to Orlandi, Bayesian perception is another, as it is the case of biased (by the frequency structure of the environmental signal) processing, akin to Gibsonian theories (that's why they call their account 'The Embedded View'; Orlandi, 2014, ch. 1).

To summarize, we can say that decouplability consists of distance and distalness. I want to make two comments on that, though. Firstly, that makes decouplability a disjunctive property. There is no specification in Orlandi's theory about what the relation between the two conditions is. They are surely not independent, as Orlandi considers them cases of abstraction. If I read them correctly, both are sufficient but not necessary conditions. However, there is significantly less clarity about whether there is specialization or overlap between distal decoupling and distance decoupling. And if the latter is the case, what is the scope and mechanism of such overlap?

Secondly, we said that Orlandi openly claims that distalness is a form of abstraction in their apparatus. They observe some similarities with Cantwell Smith's notion of 'registration,' which he equates with the system's ability to ignore details. It is the ability to control for the sources of variability and discard its byproducts. We can observe how similar these ideas are to the discussed concepts in AI research of controlling the nuisance parameters and transformational abstraction. However, it is debatable whether the information presented to the network in this way is truly

‘distant’ (or if that category applies to artificial neural nets with no perceptual systems in the traditional sense).

Detachability is not a binary switch; it is a structural capacity that interacts with the temporal demands of the task. To fully capture the difference between a mere causal trace (like an afterimage) and a genuine representation (like a memory), we must map this structural property against the temporal mode of processing. This yields a distinction between being merely 'offline' and being truly 'detached.'

5.2.1 The Temporal Dimension: Online, Offline, and Detachable

A critical distinction must be drawn between the structural property of detachability and the temporal mode of processing (online vs. offline). While often conflated in the literature[34], they are orthogonal dimensions of the informational state. Clarifying this distinction is essential to prevent the common confusion that "offline" is synonymous with "memory" or that any processing occurring in the absence of a stimulus automatically qualifies as representational. As we will demonstrate, it is possible for a system to harbor offline states that are not representations—such as afterimages—precisely because they lack the structural independence required for genuine cognition.

5.2.2 The Cognitive State Matrix

We can map representational candidates onto a matrix defined by their temporal coupling to the stimulus and their structural independence. This intersection yields four distinct zones of cognitive processing, ranging from simple reactive coping to genuine high-level cognition.

---

[34] For example, neither Gärdenfors (2005 nor Haugeland (2019) seem to differentiate and treat detachability and offline as coextensive (at least as I interpret them).

	Coupled/Non-Detachable (State is causally locked to input)	**Detachable/Structural (State stands in for input)**
ONLINE (Interaction with here-and-now)	**1. Reactive Coping** Reflexes, tropisms, optical flow	**2. Augmented perception** Emulators, predictive coding, mental rotation
OFFLINE (Interaction with there-and-then)	**3. Sensory inertia** Afterimages, priming, decay	**4. Higher-order cognition** Memory, planning, counterfactuals

Table 2: The Cognitive State Matrix

Now, let's go through all 4 parts of this cross-section one by one with pertinent examples.

Online + Coupled (The "Reflex" Zone)

In this quadrant, the system tracks the "here-and-now" via direct transduction. The cognitive state is fundamentally a relay, not a vehicle; it is causally locked to the input stream. This captures the domain of reactive coping, where the system reacts to the immediate environment using direct causal links. Examples include the Watt governor, where the angle of the arms is physically locked to the speed of the engine. Similarly, in biological systems, phototaxis describes a moth flying toward a light, and Gibsonian "pick-up" involves adjusting step length based on the current optic flow of the pavement. In all these cases, the information is processed online, but the vehicle is inseparable from the immediate causal pressure of the environment.

Online + Detachable (The "Augmented" Zone)

Here, the system engages in real-time interaction but uses a detachable model to "fill in" sensory gaps. The task remains online, but the vehicle is a simulation. This corresponds to the domain of "augmented perception" or Rick Grush's "emulator" territory, where the system uses an internal model to predict the sensory stream. A prime example is short-term object permanence: when tracking a car as it passes behind a pillar, the task of driving is online, but the car is momentarily invisible; the driver uses a detachable model to effectively "hallucinate" the car's position until it reappears. Similarly, in motor correction, the brain predicts the location of the arm before sensory feedback arrives (a delay of roughly 100 ms) to adjust movement smoothly. A more cognitive example is found in the game Tetris, where a player mentally rotates a block to see if it fits a slot while it is still falling on the screen.

Offline + Detachable (The "Cognitivist" Zone)

This quadrant represents the classic definition of high-level, genuine cognition. The stimulus is absent, and the system manipulates a stable vehicle. This includes episodic memory, such as recalling one's breakfast, or planning, such as navigating a mental map to determine a route for tomorrow. It also encompasses counterfactual reasoning, such as contemplating, "What if I hadn't sent that text?". In these cases, the representational vehicle is fully detached from the immediate environment and is subject to arbitrary manipulation by the system.

Offline + Coupled (Sensory Inertia)

This quadrant represents the "uncanny valley" of cognition. Here, the stimulus is gone, but the system retains a decaying causal trace. These states persist after the stimulus is removed but lack the structural independence to be true representations. This zone contains phenomena of sensory inertia and hysteresis, where the causal effect of the world lingers, but the system cannot control or manipulate that lingering effect. It is the "echo" of the world, not a model of it.

Consider the phenomenon of the afterimage. If one stares at a bright lightbulb and then closes their eyes, the image of the bulb persists. This state is technically "offline" because the lightbulb is off, or the eyes are closed. However, it is strictly "coupled" to the physiology of the retina. The subject cannot mentally "rotate" the afterimage, "save" it for later retrieval, or combine it with other thoughts; the system merely suffers the cooling down of photoreceptors. It is a decaying causal trace, not a portable representation.

Similarly, biological priming exhibits this offline yet coupled status. A loud noise may create a startle response, and five seconds later—when the event is clearly

"offline"—a quiet noise elicits a disproportionate response because the neurons remain "primed". The system is physically altered by the event, but this state is not a symbol for "loud noise"; it is simply a residual electrochemical facilitation. A physical analogue is damped oscillation: if one pushes a swing and walks away, the swing continues to move. The movement is "offline" relative to the push, but it is strictly coupled to the physics of the energy dissipation. The swing is not "remembering" the push; it is simply dissipating the energy.

5.3.2 The Detachability Criterion

Crucially, the quadrant of sensory inertia (offline + coupled) fails the test of representation. As illustrated by the afterimage, a state can be "offline" without being representational because the system cannot manipulate it; it merely suffers its decay. In the philosophy of information, consistent with Ramsey and Godfrey-Smith, such states fail the "Job Description Challenge". They are not true representations because they are not detachable—they are merely "laggy" causal effects.

Genuine representation requires detachability: the capacity of the vehicle to be repurposed and reactivated by the system arbitrarily, independent of the causal history that birthed it. To qualify as a representation, the system must be able to use the state in a way that is distinct from the passive dissipation of energy. Thus, I posit that while offline processing is the *function* characteristic of high-level cognition, detachability is the *structure* that enables it.

5.3 Portability

In the first chapter we have discussed the problem of how mutually communicable informational systems can be, especially when the two sides of this communication

are of different kinds. Brain-computer interfaces were an example of that. We have also considered thought experiments of informational pipelines being 'taken out' of their original setting and wondered whether they would still work in the same way. Informational pipelines were then contrasted with actual mechanical networks such as hydraulic systems, with the conclusion that there exists an important disanalogy between them. The debate on the portability of representational systems is a cognate of that.

Informally, it is a question of 'If we rip the particular representation (or a subsystem) from a bigger whole and transplant it into some other context, would it work and retain its original function and content?". In AI research, the problem is set up a little differently. It is a challenge of developing a representational scheme for a particular task and then checking if it would improve performance in another (usually quite similar) task without any additional learning and tuning (in this particular subsystem).

Take James, Rosman, and Konidaris' (2019) study on the development of egocentric space for navigation. The agent is trained to arrive at some useful representation of a toy space (a simple maze with windows) first. Then, he is moved to a different problem space (for example, one of them is a platformer-type of video game) and asked to learn how to navigate this new environment while retaining the learned representations from the first task. The authors report that using the representations of egocentric space while performing subsequent tasks is not only exploitable in new contexts but actually accelerates the learning process of a new problem. It suggests not only the possibility of inter-domain transfer of portable representations but also justifies the utility of having portable representations in the first place.

The notion of portability is also important for discussions in the philosophy of science on scientific representation. This problem is not directly relevant here, but I want to outline it very briefly, as it gives a useful intuition about what kind of property we care about here. Portable representations there are defined as concepts understandable for diverse epistemic communities because they are invariant to context (Lidskog, 2014:674). It draws heavily from Latour's (1987) work on 'writing things together' for explaining the success of cognitive artifacts in (especially modern) science - diagrams, maps, mathematical equations, statistical models, etc. They enable 'traveling' between different contexts that vary temporally, geographically, and culturally (Asdal & Jordheim, 2018). They are powerful by virtue of their transplantability.

In the philosophy of cognitive science, Cummins and Poirier define (trans)portability in terms of source independence. To illuminate the difference, they contrast them with source-dependent indicators. They use a Fodorian understanding of indicators, in which indicators are binary signals that simply register the presence or absence of their source. Had the source not been activated, the indicator would remain silent. Indicators are arbitrary but remain unchangeably bound with the source and, as such, are not portable. At the same time, they remain agnostic about the structural properties of the target.

Representations, on the other hand, work conversely, the authors argue. They do not inform about whether their target has been activated but inform that it is such-and-such. In their words:

'Indicator signals "say" their targets are present but "say" nothing *about* them; representations provide structural information about their targets but do not indicate their presence. Indicator signals say, "My target is here," while

representations say, "My target, wherever it is, is structured like so." (Cummins and Poirier, 2004:25).

Also, because the indicators are arbitrary, in principle they can indicate every possibility; they are 'chainable,' in other words, they act as transitive informers, while representations do not. If A indicates B and B indicates C, A will also reliably indicate C, on their reading. Representations, however, do not have this kind of information-preserving property[35]. 'Disciplined structural transformations of them are not going to systematically alter indicator's meanings. Susceptibility to such transformations, however, is precisely how representations earn their keep.' (op. cit.:23).

This criterion of portability faces a formidable challenge from the 'phylogenetic refinement' approach of Paul Cisek (2019). Cisek critiques what he terms the 'Cognitivist Gambit'—the assumption that higher cognition requires representations that are independent of sensorimotor modalities. Instead, he argues that the brain evolved to maintain real-time control loops, trafficking in 'pragmatic representations'—specifications for action (affordances) that are inherently 'stuck' to the motor systems they serve (Cisek, 2007). On this view, a neural state represents 'grasping vector A' rather than the objective fact 'there is a cup.' If such a state were 'ported' to a non-grasping context, it would lose its semantic content. Thus, for Cisek, the demand for portability is an evolutionary anomaly—a relic of the computer metaphor that ignores the embodied history of the nervous system.

---

[35] It does not equal saying that representations do not retain information. They do, otherwise they would be quite useless, but they do not do it transitively, as indicators do. They have other useful features that indicators lack, such as structural preservation.

However, the claim that neural representations are inherently 'stuck' to their local sensorimotor context—and thus non-portable—is fundamentally challenged by recent experimental breakthroughs that demonstrate the successful 'transplant' of these vehicles across biological and spatial boundaries. If representations were merely transient properties of a specific body-environment loop, they could not be detached and re-instantiated in a different system. Yet, this is precisely what brain-to-brain interfaces achieve. As we discussed them in Chapter 1, Section 6.2 (regarding the receiver-relativity of neural codes), the landmark study by Pais-Vieira et al. (2013) proved that motor intentions can be recorded, digitized, and transmitted between organisms across continents. The successful 'porting' of this motor plan demonstrates that the informational vehicle possesses a structural independence that survives detachment from its original biological host.

Furthermore, the manipulability of these representations—the ability to intervene on them from 'outside' the system—provides strong evidence for their status as discrete, causal vehicles rather than mere dynamical couplings. Ramirez et al. (2013) demonstrated the ability to artificially 'inception' a false memory in mice using optogenetics. By tagging the neurons associated with a specific context (Chamber A) and optically reactivating them while the mouse was being shocked in a different context (Chamber B), they created a robust fear response to Chamber A—a place where the mouse had never been shocked. As Piccinini and Thomson (2018) argue in *Neural Representations Observed*, the fact that neuroscientists can routinely observe, record, and manipulate these structures—treating them as distinct variables that can be altered or transplanted independent of the animal's immediate environment—vindicates their status as real, portable representations. If we can grab the vehicle, change its content, and put it back (or put it in another animal), it is not merely a pragmatic loop; it is a portable code.

To sum up, for their money, you can transplant representations from system to system, and they would adjust their meaning while retaining the functional role they perform. Shea (2018:224-5) takes portability to be important for the relation not between entirely different systems or agents, but between sub-personal and personal levels of human agents. He argues against portability because he thinks that portability, to be meaningful, would assume unchangeability of content. And that he thinks (in line with Cummins and Poirier) is not possible, because sub-personal states are tightly-knit with their sources. He writes:

'The contents of subpersonal representations are not things that are sitting around ready to be used by personal-level processes. It is a mistake to think of the content of a vehicle as a routinely portable property, a property that would automatically be carried around if that vehicle were deployed in a different cognitive process. The absence of a straightforward connection here means there is no simple way that personal-level contents are determined by the contents of any subpersonal representations they make use of.' (op. cit.).

However, we can see that he does seem to share their differentiation on indicators and representations in the case of sub-personal states. He thinks sub-personal states are representations, just not portable higher-up (we remember Cummins and Poirier thought the indicators were not portable).

The recent debate between Egan (2020) and Ramsey (2020) follows similar threads. Egan invokes two instructive cases. Let's imagine some brain filter within the brain matter that, from a mathematical point of view, performs Laplacian filtering. Then this subsystem is taken from the bigger system of the visual cortex and thrown in the mix elsewhere: into the auditory cortex. In the new workplace it performs the same mathematical operation and seems to be 'accepted' by the system - it performs its function well. Different embedding does not change the essential

feature of the system, which is individuated on grounds of purely mathematical content. More traditionally tokened 'cognitive content' is brushed aside as merely 'intentional gloss,' not part of theory proper (see discussion on Egan's deflationism). But methodological positions are not at stake here.

Ramsey accepts Egan's position on the purported portability of the system. He, however, disagrees about its import on the assessment of the subsystems' stability of functional role and cognitive content. When we remove the biological organ and transplant it elsewhere, it does not lose its original function, he argues.[36] But this is not the case for representational (cognitive) content.

If we take a fuel gauge, designed and implemented to monitor and inform about the amount of fuel left in the tank, and take it out and then put it inside the water tank control board, what will it represent? Even if we think that it retains its (primary) functional role, we would be hard-pressed to say it represents fuel levels. We would be similarly reluctant to say that this kind of description is superficial to both explaining the new status of the gauge and also the change of its workings.

To sum up, Ramsey says that when we have two systems that are equivalent from a mathematical point of view (they compute the same function), the importance of the 'why' question becomes more, not less, important. It is consistent with Shagrir's (2010) reading of Marr's (1978) analysis of computers performing the same

---

[36] I think he is mistaken here. He writes, 'Biological subsystems often play their primary role only by virtue of being embedded in a very specific system that brings with it the relational properties that help define that role' (Ramsey, 2020:75), and later says that the human heart would not lose its function to pump human blood even when taken out and used in an alien body to do something quite different (say, store calories). But if I undergo surgery and have a piece of quadriceps removed to be repurposed in place of torn ligaments, the sheaf of sinews clearly performs a new function. Instead of supporting the working of the big muscle, now it works in the business of keeping my foot in place. Ramsey could probably say that then it does not perform its 'primary function' (taken externistically as what the hearts usually do, qua pumps), but that does seem to beg the question to me of what constitutes 'function' in general.

operations in different contexts. When we take the same additive modules and use them first as cash registers (adding up prices of goods) and then as adders of some arbitrary coin value (defined by the same numbers minted on the coins used for purchases in the first scenario), we need the understanding of the context to fix the content. Speaking of context, we mostly mean understanding why a particular operation (and not some other) is performed. In the case of cash registers, addition is used because it happens to mirror the financial reality of how buying and selling products work in our social world, as Ramsey very poignantly observes.

To recap even more briefly: mathematical equivalence[37] affords (trans)portability from embedding to embedding but is not primary for explaining them content-wise[38].

5.4 Mirroring and Structural Resemblance

We already mentioned that good representational systems should be able to preserve and distill information about the environment. But any preservation or generalization seems moot without veridicality (or verisimilitude) baked into that. In other words, properties of the target systems have to be mirrored by the representing system in some way. This is precisely what drove Dretske in his

---

[37] Of course, the equivalence does not have to be as strict as in this imagined example. As we have seen in the tasks used for training AI agents, only some sort of acceptable level of similarity (freely thresholded for given purposes) is required

[38] Ramsey actually goes even further, saying that if one is obstinate enough to say which description of content - mathematical or cognitive—is leading and which is auxiliary, it is the other way around that Egan would want. For he says we could very well be quite knowledgeable about some particular layer performing what happens to be mathematically a Laplacian filtering, even when ignorant about this computational operation itself. As long as we understand what kind of action is performed and for what reason, we are better off than knowing which mathematical operation it instantiates. Arguably this happens quite a lot in neuroscience, when a lot of disagreement remains over which mathematical function some neurons (or types of neurons) instantiate (ReLUs, softmax, or others) even when there is good orientation about the role the particular subsystem performs.

careful attention to preserving the Xerox principle, regardless of its critiques. We will call this property 'mirroring,' as I believe 1:1 mappings are neither feasible nor desirable for effective cognizers. Mirroring seems to be a better metaphor, as it allows for certain systemic deformations—systemic in the sense that they are consistent in their transformations; one can think of magnifying mirrors here, for instance.

But importantly, mirroring needn't be iconic or imagistic. Retention of properties often emerges from systems set up arbitrarily, i.e., stipulated by mere conventions. What I mean here is that no facts of the world (how it is) are mobilized to establish the representational relation. It is purely artificial in the sense that another symbol would do the trick just as well (ceteris paribus; of course, we concede here that some coding schemes do better for certain tasks than others). Coding schemes of Morse code are easy examples, but let's consider a more everyday example: car registration plates.

In some systems (e.g., in Sweden), the assignment procedure is entirely random; a new string of symbols is assigned to any newly registered vehicle, insofar as the same string has not been assigned beforehand. But if the numbers are assigned in order from the domain available (e.g., range of x-digit numbers), then there is some mirroring between the temporal scale and the 'highness' of the number. We can, in principle, estimate how old the car is on the basis of its number. We can also do comparisons between them, so there are some structure-preserving relations between the tokens within the same scheme. Let's note here that the degree of such latent structure-mirroring is a gradable property.

To see that, we can look at some other, more sophisticated cases of car-plate systems, where even stronger retention of properties holds. For instance, within the Polish system, the cars are registered within regions (powiat), and each region

has its designated 'trademark,' two or three letters in the front of the string that designate from which particular region the car comes. Only then is the rest of the space assigned randomly. There is then a different kind of mirroring between the state of the world (geographical makeup) and the representation tokened. One can reliably say where the car (at least originally) comes from and a bit less reliably—as people move - where the owner currently resides. The mapping between the first letters is not random; they correspond to the first letters of the voivodeships, but there is still no mirroring of internal structure of geography (i.e., we cannot infer if EL is closer to ZK than it is to KR plates just by looking at them; we could, though, infer that WE should be (on average) closer to WL than to SK, as we know on this basis they belong to the same voivodeship).

Note that this is a different scenario than in the case of maps, where the internal relations within the map are deemed to be directly representational of the properties represented. The map's scale is designed and calibrated precisely to enable this kind of reading and reasoning - a stationary relation between the object and the subject. If 1 cm on the map equals 3 cm in reality, then this universally holds for the whole map. Granted, the preservation does not have to follow exactly this road. After all, the most common world maps use Mercator's projection, which famously does not retain the same ratio of representation of size for all the latitudes (that's why the northernmost and southernmost countries appear unnaturally bigger than the equatorial territories). But other properties are retained, and distortions are gradable - in the case of Mercator maps (optimized to fit the geoid into a 2D plane), even after doing this projection, it is still clear which country borders which.

Of course, the choice of representation is dictated by pragmatic considerations; they are selected for the purpose they are supposed to serve; e.g., we could care more about angular than linear distances. But what is common for all the maps is

the retention of some external properties (and discarding the others, e.g., the geomorphology for a political map) within the internal properties of the representation. In other words, there is mirroring of the ‘worldly state’ and its representational stand-in. So, we can say that whereas the mirroring is not a necessary property of stand-in relations, as it can be purely arbitrary and random (in assignment, during the execution of established protocol, it has to be rather stable to make sense), it is an important feature of some canonical types of representations, such as maps.

However, the notion that representation requires isomorphism (or even homomorphism) has been subjected to scrutiny. Pero and Suárez (2016) argue that structural mappings alone cannot account for the phenomenon of representation because they fail to handle misrepresentation adequately. They distinguish between two types of representational failure: mistargeting (picking the wrong target) and inaccuracy (representing the target imperfectly). Crucially, they identify three distinct forms of inaccuracy that a robust representational system must support:

1. Abstraction: The representation neglects some properties of the target (e.g., the billiard ball model of a gas neglects the temperature dependence of viscosity).
2. Pretence: The representation ascribes properties to the target that it does not possess (e.g., billiard balls are hard and shiny; gas molecules are not).
3. Simulation: The representation simultaneously abstracts and pretends (e.g., modeling a gas as a collection of elastic spheres involves both ignoring some gas properties and attributing some sphere properties).

This critique suggests that "mirroring" cannot be a naive identity relation. Instead, it must be understood as a second-order structural similarity. It should appear clear now that there could be various ways of adequate mirroring. Particularly, the

matching principle could be arbitrary or based on some pragmatic reasons. For many philosophers and cognitive neuroscientists, that fundamental organizational property in truly representational cognitive systems is similarity. There are ample options for similarity-based cognition on offer (Shagrir, 2012; Jacobson, 2015; Kriegestkorte and Kievit, 2013; Sachs, 2018; Gładziejewski and Miłkowski, 2017; Williams and Collings, 2017; Shea, 2018). Segundo-Ortin and Hutto (2019) call the guiding hypothesis 'Similarity-Based Cognition' (SBC). SBC occurs 'whenever an agent or system exploits the similarities that hold between two or more items—e.g., events, processes, objects, and so on.' This idea is usually cashed out in terms of structural similarity, where whole-part relations of representational vehicles correspond to whole-part relations of the targets. In other words, there is both global and local correspondence, where the arrangement of parts is also information-conveying, as long as they retain properties of the patterns of parts within the target system.

O'Brien and Opie (2004:14-15) explicate structural similarity as follows: 'One system structurally resembles another when the physical relations among the objects that comprise the first preserve some aspects of the relational organization of the objects that comprise the second.' For example, when constructing political polls, pollsters look for towns or communities in which certain crucial relations between parts (i.e., demographic relations, vote-leaning, and connections thereof) are mirroring the more complex but corresponding patterns within the entire country.

However, when we are interested in structural similarity for cognitive agents, we are presented with a more constrained case. For neural items, do not enable cognitive work to get done by possessing just any kind of similarities with their distal targets. Rather, the similarities at issue are thought to be of a second-order structural kind. Also, pitched at the neural level, is that neural states make a causal contribution to the processes that drive the behavior of cognitive systems towards

some target items in virtue of the fact that certain similarities hold between the neural states and those target items. Already invoked, Segundo-Ortin and Hutto call this "neurally based SBC" (2019).

For the purposes of this overview part of the thesis, I will not dwell further on that problem. Let me just note here that I will offer a full-fledged defense of the importance of similarity in chapter X. In fact, I will argue that SBC-type accounts are constitutive of any efficient representational systems, and they explain why certain semantically sensitive architectures emerge in a stable manner across the whole range of agents. Also, following Miłkowski (2025), I will defend the position that similarity-based structural accounts of cognition provide us the link between information and representation.

But let us not get ahead of ourselves; for now, we remain firmly within the intricacies of representationalism.

5.5 The Possibility of Misrepresentation

It was already mentioned through invoking Pero and Suarez's critique, but one crucial criterion remains. For regardless of many important properties that the account of efficient representation must possess, it can't be limited to accounting for successful cases of representation. To paraphrase a well-known phrase, 'no representation without misrepresentation.' As we have already discussed, a representational link not only is not but also cannot be rigid to perform its function. It is easy to see why, of course. For starters, it has to include the possibility of error: an adequate theory of representation must explain how a mental or pictorial representation can be mistaken. Any theory that would yield a perfect correspondence would not be very practical for our explanatory goals.

It is equally easy, perhaps, to see whether informational theories are especially endangered by struggles to explain misrepresentation. As we remember, informational theories usually explain content-bearing in causal or correlative terms. Targets cause certain occurrences of perceiving them, and that's why signs that stand for them code for these targets. Similarly, if the occurrence of a certain target (reliably!) correlates with the token T, it is enough to say that token T is about the occurrence O. The sound of the doorbell is about the person at the door because the person there either causes the sounds by activating the bell or at least these two events very reliably co-occur.

These causal-informative theories (such as Dretske's early view) are hard-pressed, then, to say how misrepresentation could emerge. After all, if the causal link is enough to fix the content, any (even aberrant or anomalous) circumstances that lead to the tokening of the sign would constitute its content. The mistaken or faulty tokening could not be interpreted as such; causal-informative theorists would have to accept the shift in the content.

Famously, this invited Fodor's (1984) critique of what he called 'Wisconsin-style semantics.' He demonstrated that the inability of these theories to disqualify errant causes leads to the Disjunction Problem. In a famous example, if a representation HORSE is learned to mean 'horse' but is then triggered by a cow in a dimly lit environment, a pure causal theory is forced to shift the representation's content to "horse OR dimly lit cow". Such a 'new' alternative is then considered a 'new' correct representation of the target—now a disjunctive class, without the ability to disambiguate between this and the 'nomic' normal 'horse means horse' case.

Fodor's response was to conjure the asymmetric dependency condition (more on it in ch. 4), which pries apart informational relations from representational ones on the basis that informational relations depend upon representational relations, but

not vice versa. In other words, gimmicky instances of triggering the representations piggyback on ones brought about during normal conditions. A counterfeit banknote produced by a forger draws its effectiveness from the fact that such-and-such pieces of paper are established as legal tender. The relation is asymmetrical: the fake depends on the real, but the real does not depend on the fake.

The challenge of solving the disjunction problem and grounding the norms of representation (what the system *should* represent vs. what it *does* represent) leads us directly to teleosemantics and forward-looking theories of content. However, exploring the specific mechanisms by which these theories fix content requires a dedicated analysis of the semantic layer itself. Therefore, we will reserve the detailed discussion of Millikan, Neander, and Buckner for Chapter 4 (Content). For now, it suffices to establish that the capacity for misrepresentation is a constitutive property of the representational vehicle, distinguishing it from a mere natural sign.

6. Conclusion

We began this chapter by asking if the concept of representation is a scientifically necessary posit or a philosophical relic. Through the lens of the indispensability problem, we saw that while simple behaviors can be explained by dynamical coupling (as the 4E theorists argue), "representation-hungry" tasks involving absence, planning, and counterfactuals require internal states that do more than just relay signals.

The I+X Strategy provides the roadmap for a naturalistic theory of representation that avoids the pitfalls of both naive symbolism and radical eliminativism. We have identified the necessary "X" factors that transform raw information (correlation) into genuine representation:

1. Decouplability: The ability to coordinate with absence and track distal features.
2. Structural Mirroring: The preservation of second-order relational structure that supports surrogative reasoning (while allowing for abstraction and pretense).
3. Misrepresentation: The capacity to be evaluated as false, grounded in a normative function.

These properties define the functional profile of a representation. But identifying the profile is only half the battle. We must also understand how these representations are physically instantiated in the brain to be computationally exploitable. A representation is not just content; it is a physical structure that must be accessed and manipulated. This leads us to the problem of format, which we turn to in the next chapter.

Chapter 3: Format

1. Why Do We Want or Need Formats?

The main motivation for mobilizing formats explanatorily is a fairly simple one: it is intuitively commonplace for differing representational vehicles to carry the same information. However, simply possessing the same informational signal (or content, if one concedes that informational states are content-bearing) may not be equally utilizable by the same processor. Computer files are the most obvious example of that. I can have the same e-book saved as .mobi or .epub, but as many readers are painfully aware, that might not equal being able to open both on the same e-reader in a legible fashion. Surely, these differences may come through differing 'supplementary' instructions within the text (e.g., how it is justified or typeset), but ultimately the same content (strings of characters) would be either unavailable (or just clunky) for a certain reader, which would be the downstream consumer in our toy example. To account for that, we need a distinction of what is represented (content) and focus on how it is represented, i.e., on the vehicle or medium of representation. Formats appear to provide the latter with being (in principle) neutral about content, which might be identical or at least equivalent.

While one might dismiss appeals to format as reflecting a lazy reliance on computer metaphors—treating mental representations as analogous to differently formatted digital files—this concern underestimates the concept's genuine explanatory utility. Format distinctions are, in fact, employed widely throughout the cognitive science literature and perform important explanatory work in contemporary research (Heinen, 2024; Pacheco-Estefan, 2024). As Cox (2014: 189) observes, decoding research on the visual system implicitly recognizes that the fundamental problem

of vision is not primarily one of information content, but rather one of format. This insight reflects a broader methodological commitment within neuroscience: the search for the brain's 'real' information—its mental currency—must attend to questions of decodability, which are inherently questions about representational format.

1.1 Informational vs. Computational Equivalence

Within computational theories of mind, the challenge of explaining representation through information requires attention to three distinct ingredients: contents, which specify what is stored; vehicles, which determine what does the storing; and formats, which address how information is stored. These three components are frequently conflated in both philosophical and scientific discourse, raising the question of whether the notion of format might be deflated or eliminated entirely. However, the main intuition motivating format distinctions resists such deflationary moves: the same information can be formatted in different ways, with significant ramifications for how that information can be processed downstream, affecting both its legibility to cognitive systems and its computational exploitability (i.e., whether it is actually usable during the subsequent processing stage).

In their influential 1987 paper "Why a Diagram is (Sometimes) Worth Ten Thousand Words," Larkin and Simon introduced the concept of informational equivalence to distinguish it from computational equivalence. Two representations are informationally equivalent if they contain the same information—that is, if any proposition that can be inferred from one representation can also be inferred from the other. Crucially, informational equivalence doesn’t require that the two representations make this information equally accessible or easy to extract. The representations carry the same content but may differ dramatically in format.

This contrasts with computational equivalence, where two representations are equivalent only if the information they contain can be extracted with similar computational efficiency. Computationally equivalent representations not only contain the same information but also support the same ease of processing. Larkin and Simon claim that while diagrammatic and sentential (i.e., linguistic) representations might be informationally equivalent—containing the same propositional content—they are typically not computationally equivalent. Diagrams can make certain information more explicit and reduce search costs by grouping together information that is used together and by using location to organize related elements. This explains why diagrams are often more useful than equivalent sentential descriptions: not because they contain different information, but because their format makes relevant information more accessible for computational operations.

This foundational distinction between informational and computational equivalence raises immediate questions about how formats are actually designed and deployed in practice. Before turning to philosophical taxonomies of format types, it is instructive to examine how the engineering community—particularly in artificial intelligence and machine learning—has grappled with the practical problem of finding the right representational structures for computational systems. Their struggles and solutions illuminate the stakes of format choice in concrete terms.

1.2 A Short Look Into Engineering. AI & Representation Learning

Let's start with AI. The first thing one observes when reading the technical literature is that representation is almost always contrasted with ‘raw data.’ Raw data is the chaotic matter on which an algorithm operates—pixels, sound waves, or sensor logs—but as such, it is rarely useful. The primary task of the AI engineer is

routinely described as 'finding the right representation' for this data. In fact, this is the central bottleneck of the field: Google's Machine Learning Crash Course (2018) explicitly teaches that feature engineering—the process of crafting these representations—is often the most time-consuming part of the ML workflow, with practitioners frequently citing that it consumes the vast majority of their project time.

At the same time, the literature uses the term with surprising breadth. On a trivial level, it can refer to simple formalisms: Roman and Arabic numerals are treated as different representations of the same object, as are Cartesian versus polar coordinates. But in the context of Deep Learning, the usage becomes more structural. Goodfellow et al. (2016:3) note that the performance of simple algorithms depends entirely on how data is presented. In their example of a system recommending Cesarean sections, the AI is not given the "raw" patient (whom it cannot see) but a structured representation consisting of specific features hand-selected by a doctor, such as the presence of a uterine scar.

This practical usage finds rigorous theoretical support in the work of John Sowa (2000), who argues that neural networks should not be viewed as alogical or non-representational alternatives to classical systems. Instead, Sowa classifies them as valid knowledge representations: they function as 'surrogates' for external entities, allowing the system to simulate or reason about the world. On this view, the neural network is simply a physical substrate capable of implementing abstract mathematical forms (specifically, procedural scripts). Thus, the 'representations' in deep learning are not metaphors; they are mathematical structures that can, in principle, be defined and described within the unifying framework of logic.

However, the modern shift toward *learning* these representations—rather than hand-crafting them—has forced engineering to look beyond just logic and into the

structure of reality itself. The struggle to make an AI "learn" a robust representation has led researchers like Irina Higgins to argue that statistical correlation is insufficient; a good representation must mirror the causal symmetry of the physical world—essentially a computational version of structural realism (Higgins et al., 2018; Schölkopf et al., 2021). Similarly, the move toward "interpretable" vector spaces reflects Peter Gärdenfors's theory of Conceptual Spaces (2000), where meaningful concepts are defined not by symbols, but by convex geometric regions . In this view, the "75% of work" reported by engineers is not merely data cleaning; it is the labor of forcing a mathematical system to rediscover the causal and geometric structure of the world.

That is why 'raw data' has to be transformed into what are called 'feature vectors' in a process called 'feature engineering.' This is not mere 'translation,' though. Features are supposed to be constructed according to 'patterns' within data. What is important is that these representations can be either created and provided manually by the programmers or learned by the algorithm itself. This process is called 'representation learning,' and as such is one of the reasons for deep learning's revolutionary prowess. Representation learning is defined by LeCun, Bengio, and Hinton (2015) as 'a set of methods that allows a machine to be fed with raw data to automatically discover the representations needed for detection or classification' (p. 436).

Goodfellow et al. (2016:5) expand on this, noting that while simple learning algorithms succeed at discovering abstract patterns, the ultimate goal of deep learning is to solve the central problem of representation: disentanglement. They argue that the most robust representations are those that separate the "underlying factors of variation" (such as rotation, translation, or occlusion) into distinct, independent variables—effectively untangling the messy correlations of the raw data.

I emphasized the key terms that, I think, are crucial tasks associated with representational jobs in artificial networks. First, representation is an abstraction-like process. Bengio says it straightforwardly: 'the central principle of machine-learning methodologies is 'the automated discovery of abstraction' (Bengio, 2013: 3). Cameron Buckner argued extensively from the philosophical side that this property of deep learning is a foundational reason for its success. He also claimed that it has important philosophical bearing as it realizes numerous conditions set on 'transformational abstraction' by British Empiricists (mostly Locke and Hume) as the primary process of human cognitive abilities.

This process of abstraction is supposed to be achieved through 'extraction' of data. Chollet (2018:28) uses the 'filtration' metaphor. According to him, the layers of the deep network act as information filters through which the more and more filtered representations are arrived at. He even calls the process 'data distillation' and the layers 'sieves' as increasingly refined filters for data, which are chained together.

Second, the pattern recognition looks for invariance - the stable elements within a dataset. Data are messy, so the learning process is conceived as looking for the sources of this messiness (noise). In more statistical terms, it looks for sources of variance to be able to correctly, e.g., classify or predict (depending on what type of network is being trained) according to this hidden structure. These sources of mess are called 'nuisance parameters.' What is encoded in nuisance parameters is what representations should overcome. When we want to classify cars reliably, we want the net to be able to abstract away from different colors, positions, tilts, occlusions, et cetera. All these properties are a nuisance. Representation is to retain the important features.

Third, because we aim to look for important features, we also want relevant features (not only invariants). If the net were to correctly ascertain cars using some

unintuitive property (which coincidentally proved to be reliable), it would be no good. The famous examples were nets that were really proficient at recognizing polar bears, but not by virtue of their anatomical sharp eye, but by looking for snowy backgrounds. That's where disentanglement is beneficial - we want the generalization to be able to discard all the spurious correlations with other properties. This condition is important not only philosophically for the interpretability of the system. It also seriously jeopardizes the representation's reusability. If the feature that the system's decision is faulty (coincidental in this case), it would in high likelihood fail when tried to be used in some other task, which we would expect to be logically consistent with learned categorization.

Lastly, and perhaps a bit trivially, the representation-learning yearns for simplicity. Representation in this paradigm is an essentially reductive process. The process of learning is routinely portrayed in the ML literature as dimensionality reduction. The system is first fed with high-dimensional data, and the point of training is to find the best possible low-dimensional representations, which approximate the initial data. Certainly the optimization conditions and benchmarks can vary (be defined according to the aims of the engineer), but generally speaking, it is always a simplifying process. It is not only a metaphorical reduction of dimensions. This is interpreted strictly as a mathematical operation.

But there are also slightly different views on representation, other than products of data distillation. Consider Mitchell's discussion of machine learning as seeking for target functions. 'Now that we have specified the ideal target function V, we must choose a representation that the learning program will use to describe the function V that it will learn' (Mitchell, 1997:8). Here, the representation is considered to be a mathematical function, not the product of the cascade of filters (which are, strictly speaking, functions themselves). As entire networks are. On a fundamental level

they are functions (albeit extremely complex sometimes) for mapping input onto output.

What is interesting is that none of these definitions consider the internal states of the network, its patterns of activations and/or weights, or connectivity schemes as doing actual representations. But this applies to hidden layers (all the middle layers between the input and output layers). For example, the final layer, which does the label mapping (in the case of classifiers), does have a clear representational interpretation. The node, whose activation is supposed to trigger the classification 'cat,' for instance, is intuitively a representation of the notion 'CAT' in the system. To sum up, there appears to be a lack of definition of what representations are, but there is significant clarity about what they are supposed to achieve in the AI community.

As for the former question, there are at least four candidates, in my opinion. ‘Representation’ could be defined as the way in which registers of raw data are encoded (i.e., format) to be used by an algorithm (broadest term); ‘representation’ could mean an extracted pattern or regularity in the data (after extraction); ‘representation’ could denote a family of functions (as per Mitchell); or ‘representation’ could stand for an internal state of the net or ensemble thereof (as the labelling nodes).

1.2.2 The Critique of "Representation" in Practice

However, this widespread and somewhat casual usage of the term "representation" in neuroscience and AI has not escaped scrutiny. Luis Favela and Edouard Machery (2023) argue that the concept of representation in the neural and psychological sciences is often used as a mere honorific, lacking rigorous definition. They contend that much of what is labeled "representation" in experimental practice—such as feature detection or correlation with stimuli—does not meet the

philosophical criteria for genuine representation. Often, what scientists call a "representation" of X is simply a neural state that correlates with X and is used by the system to perform a task. Favela and Machery urge a more restrictive usage, distinguishing between mere information processing mechanisms (which might be described dynamically or causally) and true representational mechanisms, which require satisfying stricter conditions of standing-in and decoupleability. This critique highlights the danger of allowing the "view from practice" to dissolve the conceptual boundaries of representation, reducing it to a synonym for "neural activity that matters."

The engineering practice reveals format as a practical necessity—raw data must be structured to be computationally exploitable downstream. The AI community's implicit understanding that "representation" means structured, accessible, manipulable information points toward key philosophical questions. What principles should govern format choice? Can formats be taxonomized into fundamental types? Moreover, the concern about spurious correlations and informationally greedy classifiers that emerged in our discussion of deep learning will return with greater urgency when we examine neural decoding in biological systems (Section 7). Just as deep learning systems can succeed at classification by exploiting patterns that aren't truly the relevant features—like snowy backgrounds rather than polar bear anatomy—neuroscientific decoding methods face an analogous danger. They might extract patterns from neural data that are detectable in principle but are not the formats actually used by the brain's own downstream processors. The question of format authenticity—whether we have identified the real representational structure or merely an epiphenomenal pattern accessible to powerful external tools—will prove central to assessing the biological plausibility of proposed information processing models. But first, we must survey the historical

philosophical attempts to categorize format types and understand what makes formats differ from one another.

2. Philosophical Theories of Format

2.1 Historical Accounts

The existing literature on representational formats remains predominantly philosophical in nature, yet it is both dated and surprisingly scant given the concept's importance to theories of mind. While discussions of format have become routine and well-understood within computer science, where the notion carries intuitive force, and while the concept has achieved robust standing within cognitive science, the connection between these empirical domains and the conceptual work undertaken in philosophy remains notably tenuous. This disconnect represents a significant gap that requires careful amelioration if we are to develop a comprehensive understanding of how mental representations are structured.

Philosophical debates concerning representational format have historically proceeded by drawing on intuitions about public representations, with maps serving as the paradigmatic case against which mental formats are judged. The main ambition of these philosophical investigations has been to demarcate clear distinctions between format types and to search for counterexamples that might undermine proposed taxonomies. This approach is exemplified in discussions such as Tye's (1991) analysis of the pictorial versus linguistic dimensions of the Imagery Debate, as well as in related dichotomies including digital versus analog, symbolic versus distributed, and continuous versus discrete formats. What unites these philosophical treatments is a striking paucity in the number of format types they recognize, suggesting a potential oversimplification of the representational

landscape. Some theorists, such as Raftopoulos (2020), have proposed specialized formats corresponding to different sensory modalities, though such proposals risk explanatory vacuity if they merely redesign rather than explain the phenomena in question.

Propositional accounts argue that all mental representations are language-like and amodal, meaning they are not tied to any particular sensory modality. On this view, representations have a discrete, compositional structure similar to sentences, employing abstract symbols that don't preserve the perceptual features of what they represent. These language-like structures support logical operations and inference directly, making them well-suited for reasoning and conceptual thought. Pylyshyn (1997) was a major defender of this position, famously arguing that what we experience as mental images are actually interpreted propositional descriptions—that the phenomenology of imagery masks an underlying propositional format. Similarly, Fodor's (1975) *The Language of Thought* also defended amodal, atomic, compositional representation formatted in a language-like manner.

However, propositional theories face significant challenges. They struggled to explain the rich phenomenology of mental imagery, particularly why imagistic experience feels so different from linguistic thought. They also have difficulty accounting for analog information, such as smoothly varying properties like spatial distance or color gradients, which seem more naturally captured by continuous rather than discrete representations. For many theorists, the idea that all perceptual experience reduces to propositional form remains deeply counterintuitive.

In contrast, imagery or pictorial theories claim that some mental representations are genuinely picture-like, preserving spatial and sensory properties of what they represent. These representations are depictive rather than descriptive, maintaining

geometric and topological relations in a way that mirrors perceptual experience. Stephen Kosslyn's (1980) influential research on mental rotation and image scanning provided empirical support for this view, demonstrating that people process mental images in ways that parallel how they would process actual pictures. So did Shepard's classic mental rotation experiments (1971), demonstrating analog processing, with response times increasing linearly with the degree of rotation required—precisely what one would expect if people were rotating actual image-like representations rather than manipulating propositional descriptions.

Despite this empirical support, pictorial theories face their own difficulties. The most notorious is the "interpreter problem": if mental representations are picture-like, who or what views these internal pictures? Additionally, it remains unclear how pictorial representations interface with conceptual thought and linguistic processing, and there is ongoing debate about what it even means for a representation to be "picture-like" in a non-physical, neural medium. The nature of the similarity between mental images and actual pictures remains philosophically contentious.

As such, existing theories of representational formats can be systematically grouped into three broad categories, distinguished by which conceptual component of representation they privilege in individuating formats: some theories focus on representational contents, others on the vehicles that bear those contents, and still others on the function of mapping vehicles to contents (Lee, Myers & Rabin, 2022). However, all such theories face a fundamental challenge: information processing in the brain cannot plausibly be fully captured by our commonsense ideas about icons, signs, and cartography. The neural basis of representation may require us to move beyond these familiar public analogues toward a more neurally grounded understanding of format. This difference in approach can perhaps be usefully spelled out as a difference between the philosophy of mind strand and the

philosophy of science (psychology and neuroscience, mostly) flavor. The former would be more concerned about folk psychology categories and trying to find empirical data to map them onto (or, as critics would say, pigeonhole outdated categories) in a Jamesian tradition, whereas the latter would be more inclined to find new categories to satisfactorily account for modern scientific practice (interestingly enough, proponents of this project also appeal to James; i.e., Icard, 2014).

2.2 Computational Profiles (Coelho Mollo & Vernazzani)

To escape the traditional debates on format that are seemingly stuck, verbal, or discontinued, current authors suggest it is not productive to follow this path and try to sidestep it. To give an example, in an important recent paper, Coelho Mollo and Vernazzani (2023) propose a radical reconceptualization of representational format that aims to move beyond the traditional philosophical debates rooted in intuitions about public representations like maps, sentences, and icons. Their central contention is that these familiar analogies—while intuitively appealing—have ultimately proven misleading when applied to understanding how biological and artificial cognitive systems actually process information. The brain's representational formats, they argue, cannot plausibly be fully captured by our commonsense ideas about icons, signs, and cartography. Instead, they propose that we should start from scratch to develop a framework that is genuinely responsive to empirical work in neuroscience and cognitive science, one that can account for the actual computational architecture of cognitive systems rather than imposing pre-theoretical intuitions derived from external symbolic artifacts. They openly declare, "Our purposes here are exclusively constructive—namely, to detail and defend a version of the vehicle-based approach motivated and shaped by the computationalist framework in mainstream cognitive science—and aimed at

producing a notion of representational format that can be useful and fruitful for cognitive science and AI research" (Coelho Mollo & Vernazzani, 2023: 5).

Their positive proposal centers on the concept of computational profiles. According to their first thesis, representational formats simply are the computational profiles of representational subsystems within cognitive systems, whether biological or artificial. But what exactly is a computational profile? Their second thesis provides the answer: computational profiles are individuated by the inner and outer constraints of computational vehicles—that is, by the values those vehicles can take and by the mutual constraints among them. To make this abstract characterization more concrete, Coelho Mollo and Vernazzani offer a striking analogy: think of action figures. The parts of an action figure that can be moved constitute, with considerable stretch of terminology, the "vehicles" of the system. The number of relative positions these moving parts can occupy and the relations between variations over them—for instance, whether moving one leg causes the other leg to move as well or whether it affects an arm instead—constitute the inner and outer constraints, respectively.

Applying this framework to cognitive representation, we can ask parallel questions about computational vehicles: How many parts of the vehicle can be computationally "wiggled," and what values can they take? How does wiggling one part affect the possible values of another part? And crucially, how does wiggling the values of one vehicle affect (or fail to affect) other computational vehicles—that is, does changing the values taken by one vehicle systematically affect the values of others? The answers to these questions, they argue, exhaust what it means for a representational system to have a particular format. Format is thus not a matter of resemblance to external symbolic systems, nor is it fundamentally about pictorial versus propositional structure. Rather, it concerns the computational degrees of

freedom available within a representational system and the dependency relations among those degrees of freedom.

This approach is designed to satisfy what Coelho Mollo and Vernazzani identify as the key explanatory desiderata for any adequate theory of representational format. First, such a theory should enable transformation-based explanations, helping to explain the workings and behavioral effects of cognitive states and processes in terms of the specific kinds of transformation or manipulation available to and performed over such states and processes. Second, it should support efficiency-based explanations, clarifying why certain cognitive states and processes are more or less adequate for specific tasks in terms of certain sets of transformations or manipulations being more efficient, powerful, less computationally taxing, or temporally advantageous than alternatives. Finally, and perhaps most ambitiously, a theory of representational formats should exhibit epistemic fruitfulness—it should open up new research questions, suggest novel experiments, and provide a productive framework for scientific investigation rather than merely redescribing phenomena we already understand.

The computational profiles framework provides a promising reconceptualization of format grounded in computational architecture rather than intuitions about public representations. However, to fully appreciate how formats constrain and enable cognitive processing, we need operational criteria for when information is genuinely accessible to a system. It is one thing to say that a format determines "computational degrees of freedom" and "dependency relations"; it is another to specify precisely when information in a given format is explicit, immediate, or exploitable by downstream processes. This brings us to two crucial frameworks that address the accessibility question directly: procedural explicitness and the structure of immediate availability.

3. When Is Information Explicit?

3.1 Procedural Explicitness

The discussions above highlight that the primary function of format is to render specific information computationally accessible. But what does it mean for information to be "explicit" or "accessible"? David Kirsh, in his 1992 paper "When is Information Explicitly Represented?" addresses a question of implicit/explicit information boundary differently than Kulvicki. In short terms, we can say that, according to him, information is explicitly presented whenever it does not require any additional intermediary processing step between the readout. Particularly, no computation is compulsory for mediating the access to such a piece of explicit information.

Actually, the starting intuition in the paper states that computation itself is ‘a process of making explicit information that was implicit’ (Kirsh, 1990:307). For example, when computing 5 as the solution to $\sqrt[3]{125}$, we move from a description that is not explicitly about 5 to one that is. We are drawing out numerical consequences, extracting information implicit in the problem statement. To do so, we require an additional computational step to achieve that. Intuitively, if an agent did not know what the cube root of 125 is, this piece of information is inaccessible to it as its solution. Only after performing the additional step does information become explicitly presented and ready to be used in such a format.

Hence, Kirsh develops the stance that we can call procedural, in contrast to the classical syntactic account defended by Fodor and Pylyshyn, who argue that explicitness requires the distinct syntactic physical presence of representations. In other words, it holds whenever there is a dedicated symbol or structure that carries that specific piece of information. Accessibility to the system is irrelevant; what matters is whether there’s a discrete physical token (Kirsh, 1992:343-4). It is

important to note here that the procedural view states that explicitness is always relative to some procedure for reading out the information. This procedure exploits how a code is structured. This is crucial, as there's no absolute fact about explicitness independent of what computational operations are available to extract and use the information.

But this property is gradable. The degree to which information is encoded explicitly in a representation is related to the computational cost of recovering or using the information (Kirsh, 1992:345-6). In fact, knowledge that is implicit in a system need not be represented at all (even implicitly) if the cost of recovering it is prohibitive. Let's bear it in mind as it will return with great importance in the context of informationally greedy (and thus implausible) classifiers of neural data (that purport to achieve neural decoding).

Additionally, Kirsh argues we shouldn't let ourselves be "too quickly abused by sentential and linguistic requirements when we investigate mental explicitness." Linguistic explicitness (found in words and sentences) is just one form, grounded on specific physical vehicles. Informational explicitness, however, need not mirror linguistic structure. What's more important is that information is usable (exploitable/utilizable/deployable) by the system across various circumstances. Andy Clark (1992), building on that, calls for 'a variety of modes of use', which seems to put this point quite nicely.

3.2 Easy Access

More fundamental than the usefulness of a representation is the looser but more general condition of accessibility: a representation must render its information accessible to be of any value whatsoever. This condition states that correct formatting licenses the utilizability of information—it is the necessary bridge between mere presence and actual deployment. Information could be present in the

data, whether in neural signals, raw measurements, or unstructured datasets, yet remain unusable either in principle or in practice if not appropriately formatted. The relationship between a representation and the information it carries is not automatic; poorly formatted representations create an extraction problem that can be computationally or cognitively prohibitive. As Farquhar and Farquhar (1891:55) memorably put it, getting information from numerical tables is like extracting sunlight from a cucumber—the content may technically be there, but the format renders it practically inaccessible. Legibility, then, is not merely a convenience but a constitutive condition for information's usability: without the right representational structure, information remains inert, unable to enter into the cognitive or computational processes for which it was ostensibly gathered.

To really put this point across, let’s consider two examples from the literature. First, let’s reminisce and float back to the jolly ol’ times of the Clinton administration and consider The Invisible President:

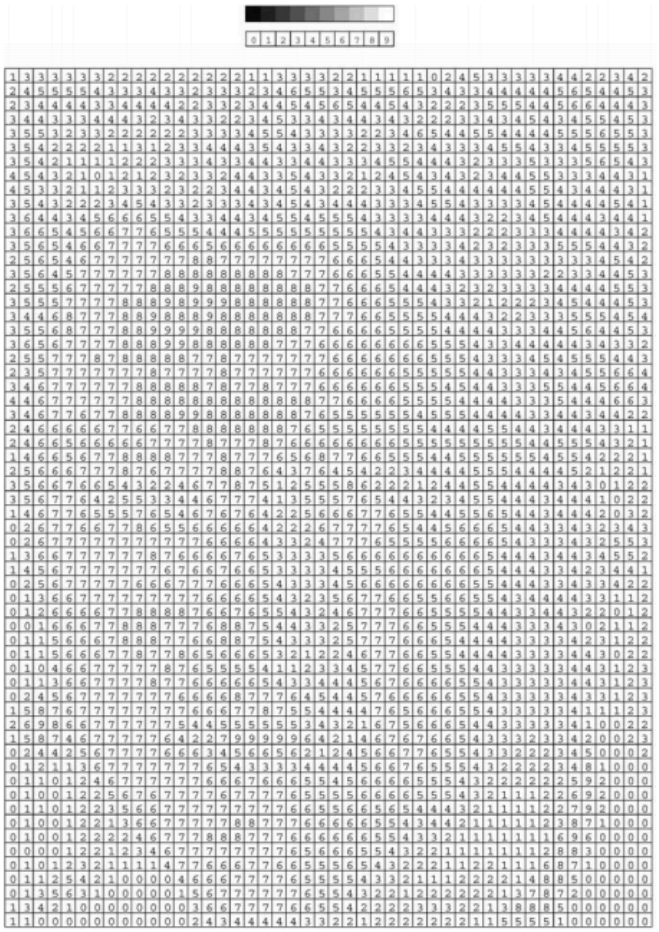

Figure 5: The Invisible President

Information is in place, but we can all agree that provided we are not some sort of savants in cucumber light extraction, it's not obvious to us what's being represented here. Sure, we can perhaps spot some tendencies if we really zoom in and put great effort - we could see some edge-like patterns of numbers of the same or similar values (5s to 7s). We could speculate that probably these might code for pixels, and these lines would be edges in the picture. But in principle, this chunk of structured data is not legible to us directly.

Now if we substitute the numerical values with corresponding grayscale pixel intensities, we arrive at this picture:

Figure 6: The Visible President; example from Hyvaerinen et al.,2009:4-5)

Now, the information becomes immediately accessible and available to us, easy to interpret. It's a slightly pixelated and entirely grey but perfectly fine portrait of Bill Clinton. He was hiding in plain sight (again, one might say). But the real lesson here is that format is supremely important for epistemic success. Even though two information packages were equivalent in regard to both outer constraints (e.g., degrees of freedom) and internal structure (1:1 values correspondence), they varied drastically in terms of their exploitability.

3.3. Floors and Ceilings

[Invisible President] is a neat example, but it's more of an intuition pump, really. A toy example, if one may. We can do better and search for immediacy, access that

could bring us closer to understanding what renders information exploitable/utilizable.

John Kulvicki's 2010 paper "Knowing with Images: Medium and Message" develops a theory of immediacy (or *immediate availability,* used interchangeably) to explain why images and graphs are epistemically valuable compared to descriptions and lists. Central to this theory are the concepts of syntactic salience and semantic salience, which work together with extractability to determine when information is immediately available in a representation.

Information is extractable when there is a nonsemantic feature of the representation that carries that specific piece of information and nothing more. For example, in a temperature graph, a red region indicates stormy weather of a certain intensity—the color red alone carries this information without saying anything about location (which is carried by spatial position). Extractability is about how nonsemantic features are responsible for content.

For information to be immediately available, the properties that carry that information must be perceptually salient—they must stand out to the viewer. Kulvicki calls this "syntactic salience" because it concerns how consumers relate to the representation's nonsemantic features, not its content.

For example, if temperatures in the 98-102°C range are represented with shades of red while other temperatures are green, the relevant features stand out perceptually. But if all temperatures are represented with slightly different shades of red, syntactic salience is lost because we're poor at discriminating between similar shades. The information is still extractable, but the features carrying it aren't perceptually prominent.

Even if information is extractable and syntactically salient, it must also be easy to learn how the perceptually salient properties correlate with features of what's represented. This is semantic salience—the plan of correlation must be graspable. For instance, mapping temperature ranges from lowest to highest onto colors from darkest to brightest creates semantic salience because the isomorphism is intuitive. But arbitrarily assigning the middle color to the highest temperature would destroy semantic salience, making the representation difficult to interpret despite syntactic salience[39].

Kulvicki's key insight is that images don't just make their determinate content immediately available—they make information across many levels of abstraction immediately available.

Consider a temperature dataset. A list of numerical triples (x, y, T) presents determinate information immediately, but accessing abstract information (like "the corners have temperature extremes") requires first decoding the specific values, then inferring the abstract pattern. The list follows a "decode first, ask questions later" approach.

By contrast, an image using brightness to represent temperature allows you to perceive abstract patterns directly—you can see that corners are darkest/lightest without first extracting specific temperature values. Perceptual abstractions over the image's features (like "this region is darker than that one") directly correspond to abstractions over the content (like "this area is colder than that one").

---

[39] This concept of 'semantic salience'—the intuitive mapping between vehicle and content—provides the epistemic counterpart to the efficiency gains identified by Mann (2024; as mentioned in a footnote in section 4.7.1 of ch. 1). Just as Mann argues that the 'right' structural format (e.g., a map) reduces the computational complexity of extracting relational data, Kulvicki argues that high semantic salience renders that same data 'immediately available.' In both accounts, the structural match between the format and the target is what allows the system to bypass expensive decoding steps.

Kulvicki introduces the metaphor of floors (the most determinate information carried) and ceilings (the highest level of abstraction immediately available). Images have many "steps" between floor and ceiling—they make information at multiple levels of abstraction syntactically and semantically salient simultaneously. Descriptions and lists typically don't provide this graduated access.

The reason images carry special epistemic weight is that interpreting them mimics perceptual investigation itself. We perceive properties at multiple levels of abstraction (both the determinate dark green and the abstract "green" or "dark"), and images preserve this structure. When we visually interrogate an image, we're using the same perceptual resources we use to investigate the world directly.

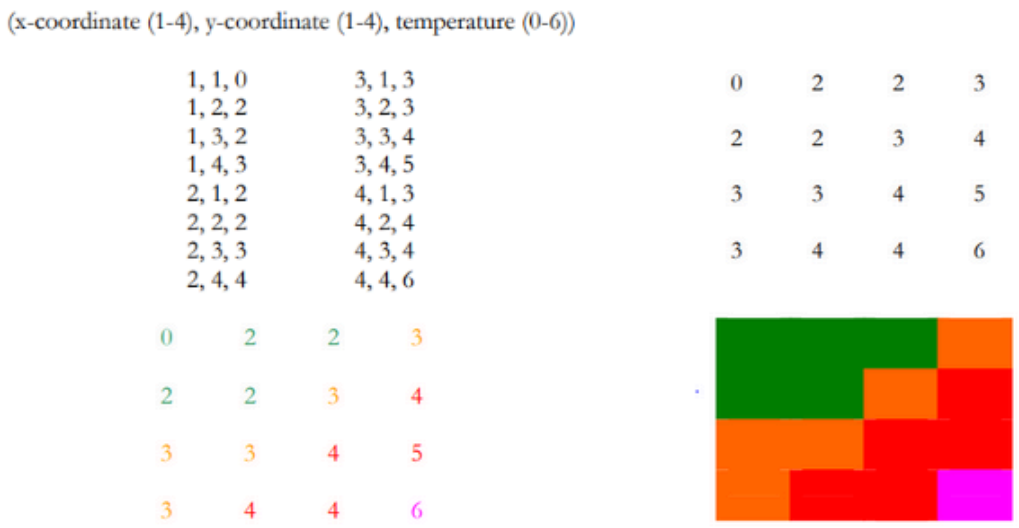

Figure 7: Differences in immediacy based on syntactic and semantic salience of equivalent (or isomorphic) information packages (example from Kulvicki [2010])

This explains why we feel "reliably and intimately connected" with images' contents—not necessarily because images are more accurate, but because our grasp of their content is obtained through perception-like means we trust from everyday experience.

We can see that Kirsh's criteria for explicitness closely match Kulvicki's. Ease of use corresponds to Kirsh's emphasis on read-out procedures and how code structure determines what information can be recovered, while semantic/syntactic distinction likewise undermine the primacy of linguistic form. But most importantly, both are fundamentally concerned with when information is formatted appropriately for use rather than merely present in principle.

Kulvicki's framework gives us the conceptual tools to analyze when information is immediately available across multiple levels of abstraction. Kirsh's procedural account clarifies what it means for information to be explicit relative to a system's computational capacities. Armed with these operational criteria for assessing format adequacy, we can now turn to empirical cases where format choice demonstrably affects performance. The philosophical distinctions discussed above might seem abstract, but they have immediate consequences for empirical research. If representational format is a real variable in cognitive systems—not merely a convenient descriptive gloss—we should be able to find cases where different formats, despite being informationally equivalent, lead to measurable differences in performance, error patterns, or neural implementation.

4. Empirical Cases: When Format Makes a Difference

Now we are primed to see it play out as a real, causally active, and explanatorily crucial posit in empirical practice. If representational format is a real variable in the cognitive system, we should be able to find cases where different formats—despite being informationally equivalent—lead to measurable differences in performance or error patterns.

4.1 Tic-tac-toe up to 15

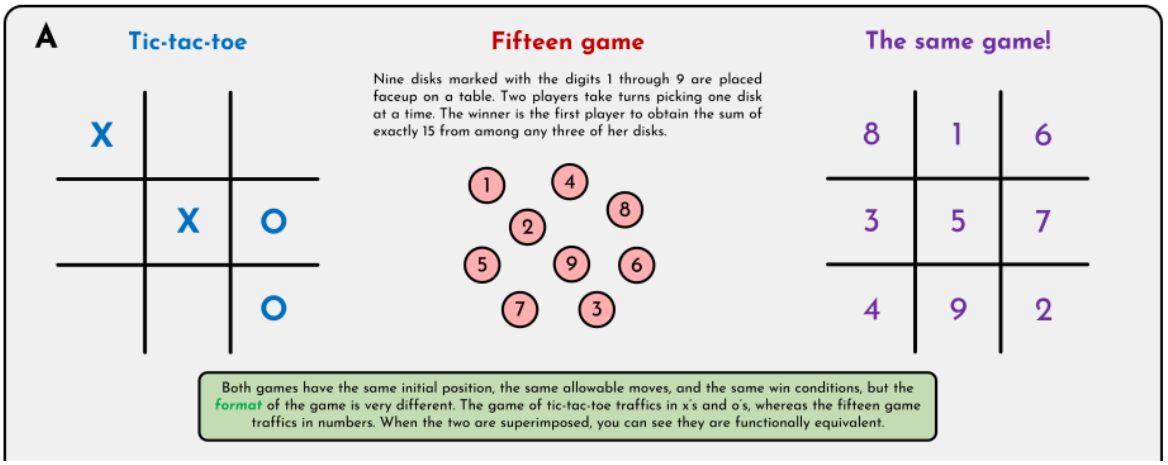

Figure 8: Tic-tac-toe vs. the Fifteen Game

Thus, we now turn to another example, this time from Yousif (2022), which presents us with two little games to play. The first one is a classic tic-tac-toe, which requires no explanation. The second game is also simple - players have 9 tokens numbered from 1 to 9 to choose from. They take turns selecting the disks, and the goal is to amass the sum of exactly 15 after 3 rounds. So here we are presented with two scenarios, which, for a change, are both quite immediately understandable. However, it is probably not immediately obvious (even to subjects that realize it quickly) that these two games are equivalent. They share initial conditions, moves that are allowed, win conditions, and its goal. However, they traffic in different formats through the game, which sets them apart not only for interpretation but plausibly for players' abilities, too. It is not difficult to imagine that a person quite well-acquainted with tic-tac-toe in a way that she would never lose a game and would struggle with not losing the second game (if a player were, for instance, a child in the process of mastering simple arithmetic).

We can see clearly now that informationally equivalent scenarios could lead to different epistemic performance. But these differences are not only hypothetical and amount to demonstrating the difference in principle. Much more importantly, format serves a critical explanatory role in prying apart rivaling hypotheses in cognitive psychology about the mechanisms and strategies employed in accomplishing certain tasks.

4.2 Cartesian vs. Polar Coordinates

Sami Yousif uses the distinction between polar and Cartesian coordinate systems as a key example of how different representational formats can encode the same spatial information with equal precision yet differ fundamentally in *how* they structure that information—and consequently in what computational operations they afford.

As Yousif explains, both Cartesian coordinates (using x and y dimensions) and polar coordinates (using angle θ and distance d) can specify any location in two-dimensional space with equal precision and efficiency. The systems are informationally equivalent: "Both Cartesian coordinates and polar coordinates capture the same level of information. Furthermore, they specify locations with the same level of precision. Both Cartesian coordinates and polar coordinates are capable of accurately describing any two-dimensional relation." What differentiates them is purely their format—how they organize the same information—and consequently what functional goals they serve.

Yousif and Keil (2021b) developed an ingenious method to determine which coordinate system people actually use by analyzing the structure of errors in spatial localization tasks. The logic is elegant: if the mind represents locations using truly independent, orthogonal dimensions (as an efficient two-dimensional format should), then errors in those dimensions should be uncorrelated. If you make an

error in one dimension, it shouldn't systematically influence your error in the other dimension.

Across multiple paradigms, they found that Cartesian errors were consistently correlated (errors in x predicted errors in y), whereas polar errors were consistently uncorrelated (errors in angle didn't predict errors in distance). This pattern strongly suggests that observers relied on polar coordinates by default to complete these visual localization tasks. The uncorrelated polar errors indicate that angle and distance are represented as independent dimensions, which is precisely what you'd expect if polar coordinates constitute the underlying format.

Critically, Yousif found that the representational format wasn't fixed. When observers performed localization tasks over a visible grid, their error patterns shifted: Cartesian errors became uncorrelated, while polar errors became correlated. This suggests the visual system can flexibly switch formats depending on environmental structure. In unstructured environments, polar (or graph-like) formats may be used by default; in structured, grid-like environments (perhaps mimicking city layouts), Cartesian formats become more useful. This flexibility hints that the mind may support both types of representations simultaneously, calling on one or the other depending on task demands—precisely the kind of redundant formatting Yousif argues facilitates flexible spatial behavior.

4.3 Ants on Stilts

Yet an even more spectacular example comes from animal cognition. In Wittlinger et al.'s (2006) study, researchers tried to empirically gain access into the strategies employed by the desert ant (*Cataglyphis fortis*) during navigation, namely foraging trips from the mound and back. This case study exemplifies what Yousif calls the "holy grail" of format research: achieving near-direct insight into the underlying structure of mental representations through simple, elegant behavioral

interventions. These ants navigate home along straight paths even after circuitous foraging routes by relying on an apparent pedometer tracking step counts combined with directional sense. This two-place vector effectively implements a polar coordinate system—storing distance (as steps) and direction (as angle)—which suffices to create a vector between any two points in space. The critical evidence comes from Wittlinger and colleagues' manipulation: placing stilts of different lengths on ants' legs caused systematic errors, with ants overshooting their nest when legs were lengthened and undershooting when legs were shortened. This demonstrates that the format is genuinely polar rather than Cartesian, as errors manifest specifically in the distance dimension (step count) while directional information remains intact.

The power of examining error patterns extends beyond insects to human spatial cognition. As Yousif demonstrates, if errors in both variables of a coordinate system are correlated, this suggests those dimensions are not truly independent and therefore unlikely to represent the actual formatting of spatial information. An efficient two-dimensional format should comprise orthogonal, independent dimensions, and errors in such a format should likewise be independent. When Yousif found uncorrelated errors specifically in polar coordinates (angle and distance) but correlated errors in Cartesian coordinates (x and y), this provided compelling evidence that human visual localization relies on polar formatting by default. The error-correlation method thus offers a window into representational structure: systematic patterns in what goes wrong reveal what's actually being stored and how.

However, while such behavioral interventions provide the gold standard for investigating format in simpler organisms, humans present greater complexity. Unlike desert ants with their narrow behavioral repertoire and easily disrupted navigation, humans exhibit flexible spatial behavior across virtually any terrain and

can effortlessly translate information across scales and modalities. This flexibility likely requires multiple coexisting navigational systems rather than a single, specialized format. Evidence increasingly suggests that humans maintain both Euclidean/Cartesian and graph-like/polar representations simultaneously, calling upon different formats depending on task demands—whether navigating directly from point A to point B (favoring polar), representing full city layouts (favoring Cartesian), or communicating directions verbally (favoring propositional formats). However, even the "holy grail" of direct format access remains valuable but limited when trying to extrapolate the results to more sophisticated systems. We can successfully decide between rivaling hypotheses for specific tasks or classes of tasks, but we must be wary that human cognition (especially human) is fundamentally pluralistic, supporting redundant formats that enable the cognitive flexibility absent in simpler organisms. In fact, this redundancy and existence of multiple parallel processing systems (scil. systems using different formats) is advocated by Yousif, Keil, and others. He opines, “However, here one might fruitfully apply this notion that representations need not be reduced to a single format. There is, after all, work that supports the role of principal axes (Marr & Nishihara, 1978; Sturz et al., 2017). As with location representations, the mind may use distinct formats in different ways depending on the task. For instance, the medial axis may be used for object recognition, but the principal axis is used to evaluate rough size (e.g., whether a certain piece of furniture will fit in a certain space). These same insights can be applied to the study of other spatial properties, such as area (Corbett & Oriet, 2011; volume (Teghtsoonian, 1965), and orientation (Appelle, 1972). In all of these cases, it may be useful to consider the possibility that any piece of information may be formatted in multiple ways simultaneously.” (Yousif, 2022:13)

Of course this idea is not new and was defended most famously by Allan Paivio under the banner of dual-coding theory. Dual coding theory posits that any given piece of information can be formatted both visually and verbally, representing an early and influential recognition that the mind might maintain multiple representational formats simultaneously. On this view, the concept "chess," for instance, may have both a visual form—perhaps an image of a chessboard with its distinctive black and white squares and recognizable pieces—and a verbal form consisting of explicit, language-like knowledge about the organization of the board and where each piece belongs. This dual representation serves multiple functional purposes. Most obviously, having information encoded in two different formats may reduce the likelihood of forgetting: if one trace degrades, the other may remain accessible. Furthermore, different formats may be better suited to different cognitive tasks. The visual representation might support rapid pattern recognition during gameplay, while the verbal representation might facilitate explaining strategies or teaching the game to others.

The adaptive value of such redundancy becomes particularly clear when considering cases of selective impairment. If information exists in only a single format and that representational system is damaged—whether through ordinary forgetting or more serious brain trauma—the information is lost entirely. But if the same information is stored redundantly across multiple formats, damage to one system need not eliminate access to that knowledge. A patient who suffers damage to visual processing areas might lose the ability to visualize a chessboard but retain verbal knowledge of chess rules and strategies. Conversely, damage to language areas might impair verbal description while preserving visual recognition. This architectural redundancy thus serves as a crucial backup system, enhancing the robustness of memory and cognition more broadly. The principle extends beyond the visual-verbal distinction: redundant formatting may characterize many

domains of cognition, with the same information simultaneously encoded in multiple ways to support the diverse and flexible behaviors that characterize human intelligence.

Dual coding theory also anticipates contemporary debates about format redundancy in spatial cognition. Just as information can be coded both visually and verbally, spatial information might be coded in both Cartesian and polar coordinates or in both map-like and graph-like formats. The key insight is that these seemingly redundant representations are not computationally wasteful but rather enable cognitive flexibility. Unlike simpler organisms or specialized computational systems designed to perform specific tasks optimally, human minds must navigate innumerable situations and solve problems that cannot be anticipated in advance. Having the same information available in multiple formats means that whatever the task demands, an appropriate representation is likely to be available. When navigating directly between two points, a polar representation specifying angle and distance proves most efficient. When integrating new locations into an existing cognitive map of a city, a Cartesian representation may better support the addition of new coordinates. When giving verbal directions, a propositional or landmark-based representation becomes most useful. Rather than constantly translating information from one privileged format into task-appropriate formats—a process that would be computationally expensive and time-consuming—the mind may simply maintain multiple formats in parallel, ready to be deployed as needed. This redundancy, far from being inefficient, may be precisely what enables the remarkable flexibility and adaptability of human spatial cognition.

5. Digital, Analog or Generic (Sui Generis) Neural Representations?

The worry about differing formats is continued via the distinction between analog and symbolic (or digital) representation and computation, is a foundational issue across philosophy of mind, cognitive science, and computer science (Piccinini, 2022; 2025). While examples like mercury thermometers and photographs are generally considered analog, and digital clocks and binary data are considered symbolic, controversy exists over precisely what constitutes the "mark of the analog" (Lee et al., 2023). Two major philosophical accounts—the Rulebound Structure Theory and Corey Maley's view—provide competing frameworks for analyzing this distinction and its implications for both representation and computation.

5.1 Rulebound Structure Theory

Lee, Myers, and Rabin (2023) develop the rulebound structure theory, arguing that the mark of the analog is located primarily in a representational system's interpretation function, rather than solely in its vehicles (the representers) or contents (the represented). Every system contains vehicles, contents, and an interpretation function that maps the former to the latter. The authors contend that focusing on the interpretation function captures what many previous theorists intended, even when they employed "vehicular rhetoric". The central claim of this theory is that analog systems utilize interpretive rules that map syntactic structural features onto semantic structural features. This perspective aligns with the broader structural approach, which assumes that analog representation involves a correspondence between structures.

To capture the reality that analogousness comes in degrees, the theory proposes three independent, degree-theoretic measures. First, analog mirroring measures the extent to which structure corresponds to structure, determining how closely a

system adheres to the isomorphism criterion. Mirroring is quantified as the ratio of the size of the maximal subsystem that satisfies isomorphism (the "mirrored subsystem") to the size of the overall system. A system must possess at least some mirroring to be considered analog, positioning mirroring as a "gatekeeper". Second, analog purity measures the proportion of the representational work achieved through structure correspondence. A system like the City Populations map, which uses letter size (structural feature) and letter character (nominal feature) to represent populations and cities, respectively, may achieve maximum mirroring but have relatively low purity because much of the difference between elements is nominal rather than structural. Conversely, a system like the mercury thermometer, where temperature is directly related to mercury height, achieves maximum purity. Third, analog mass measures the absolute amount of representational work done by structure correspondence.

The rulebound structure theory improves upon alternative structural accounts. For instance, density theories—which claim a system is analog if it is syntactically and semantically dense—are rejected because they are both too restrictive and too permissive. Similarly, magnitude mirroring theories are found wanting because they restrict analogousness to properties that are magnitudes. The rulebound structure theory can account for systems that use non-magnitudes (like hue) and can also differentiate degrees of analogousness in systems involving magnitudes, which magnitude mirroring theories treat as "all-or-nothing".

5.2 Maley's Account

Ironically, in the case of intuitions about analog computation, the prevailing idea that analog computation is essentially about continuity is contradicted by historical examples of discrete analog computers (Maley, 2018a 2023). Maley's view characterizes analog computation by focusing on a specific type of analog

representation that accommodates both continuous and discontinuous systems. According to this view, analog representation is the representation of magnitudes by magnitudes, asserting that structure preservation—or monotonic covariation—is the essential feature. Maley's account states that a representation R of quantity Q is analog if a representational property of R specifies Q, and the amount of P is a monotonic function f of Q, which is a homomorphism from Q to P. This characterization, emphasizing monotonic covariation, covers systems like continuous electronic differential analyzers as well as discontinuous electronic switching circuits.

The philosophy of computation often grapples with whether representation is a necessary component of computation. Maley's theory aligns with the orthodoxy that representation is crucial. In contrast, Piccinini's mechanistic account defines physical computation as the functional processing of medium-independent vehicles in accordance with a rule, explicitly excluding representation as a necessary condition. However, the mechanistic account faces criticism regarding its reliance on the concept of "medium independence," which is essential for distinguishing true computing systems from non-computational mechanisms. Maley (forthcoming) argues that the mechanistic theory encounters a dilemma: if medium-independence is inferred from the rules of the mechanism (the Rules-First interpretation), it counts too many simple artifacts—like cylinder locks, torque wrenches, and sprinklers—as computational. If medium-independence is provided by definition (the MI-First interpretation), the account cannot objectively determine whether natural systems like neural systems compute. This critique suggests the need to return to coupling the mechanistic framework with a semantic or representational account of computation.

The nature of analog representation, particularly its reliance on physical properties, has significant implications for how neurocognitive systems are analyzed,

especially regarding Marr's famous levels of analysis. Unlike digital representation, which represents numbers via discrete digits and abstracts away from its physical implementation (making it medium-independent), analog representation is inherently medium-dependent because it directly utilizes physical magnitudes. The medium-dependence of analog representation leads to the argument that Marr's three levels of analysis collapse into a two-level framework for analog computational systems. For analog systems, the representational and implementational merge into a single level that is simultaneously representational/algorithmic and implementational, because the physical substrate fundamentally dictates the representational format. In neuroscience, where specific physical properties of neural firing (such as firing rate or spike density) monotonically covary with stimulus properties (such as muscle stretch or temperature), these processes are viewed as analog representations. If neural processes are viewed as computational, they are often best understood as analog computation within this collapsed framework.

The rulebound structure theory also provides a framework for explaining the trade-offs between analog and symbolic systems—namely fecundity and flexibility. Analog systems are highly semantically fecund, meaning they offer high expressive power relative to the complexity of their interpretation function. This benefit stems from analog systems exploiting the pre-existing structure of their vehicles and contents. A single, simple rule (like "darker shades represent greater densities") can generate a vast, sometimes infinite, number of contents, resulting in significant interpretive efficiency. However, analog systems are generally less semantically flexible than symbolic systems. This manifests in two ways. First, regarding modularity: symbolic systems are more modular, using separate rules for content assignment, making it easier to alter or add specific assignments without affecting the rest of the system. Analog systems, conversely, are often organized such that

changing one assignment requires massive alteration of many others. Second, regarding generality: analog systems are limited in what they can represent because the representational target must possess exploitable structure, and analogous vehicles must be available to do the representing. Symbolic representation imposes no such requirement.

5.3 Piccinini's Generic Computation

A third, rival view was offered by Piccinini (2020)[40], who argues that neural computation employs sui generis (generic) types of computation that are distinct from both traditional digital and strict analog computation. According to this framework, physical computation in the generic sense is defined as the functional manipulation of medium-independent vehicles in accordance with a rule. Since typical neural signals fall outside the strict definitions of digital computation (lacking strings of digits) and analog computation (lacking truly continuous variables for integration), neural computations must be assessed as sui generis.

The claim that neural processes are computations in this generic sense rests on two premises related to their vehicles. First, medium independence: the functionally relevant aspects of neural processes—primarily spike rates and spike timing—are similar throughout the nervous system, irrespective of the physical properties of the stimuli they may encode. These features (timing and rate) can be specified abstractly and can theoretically be realized by different physical media, such as silicon-based circuits. Processes defined over these multiply realizable inputs and

[40] This sui generis framework is rooted in Piccinini's earlier collaborative work with Bahar (Piccinini & Bahar, 2013), which precedes both the accounts presented in sections 5.1 and 5.2. In that paper, they defended "a weak version of computationalism—neural processes are computations in the generic sense" and argued on empirical grounds that neural computation should be classified as sui generis, since typical neural signals like spike trains are neither purely continuous (as required for analog computation) nor strings of digits (as required for digital computation). This 2013 analysis established the groundwork for Piccinini's (2020) fuller development of the generic computation framework.

outputs are thus considered medium independent. Second, rule governance: neurocognitive systems process these medium-independent vehicles (spike trains) in a rule-governed way, meaning their manipulation is not random. Since neurocognitive processes satisfy the criteria of manipulating medium-independent vehicles in accordance with a rule, they qualify as computations in the generic sense. This generic definition includes digital computation, analog computation, and other unconventional kinds of computation, such as quantum computation, while avoiding pancomputationalism.

So perhaps we should go generic? While neural processes are considered computation generally, proponents of this view argue they are specifically neither digital nor analog. Neural computation is classified as sui generis (a kind of its own) because the physical structure of neural signals, such as spike trains, deviates significantly from the essential requirements of the other two major computational types.

The arguments against neural computation being digital include several considerations. Although spikes are all-or-none events, current neuroscientific evidence suggests that the functionally significant variables for processing neural signals are properties of spike trains, such as firing rates and spike timing—spikes themselves are not standard digits. Moreover, there is no evidence that pluralities of spikes (whether from a single neuron as spike trains or synchronous spikes from multiple neurons) can be unambiguously treated as strings of digits, where the position and identity of each element is clear. Without strings and operations defined over strings, digital computation is not present. The original proposal for the digital computational theory of cognition was the McCulloch and Pitts cornerstone paper (1943), based on the all-or-none nature of the action potential. But contemporary, realistic mathematical models of neural processes show that the

explanatory role of digital computability theory is limited in current scientific practice.

The arguments against neural computation being analog in the strict sense are equally compelling. Analog computation, in the strict mathematical sense, involves manipulating variables that can be continuous, varying over real time, and whose values must be integrated by the system. However, typical neural signals are described as graded like continuous signals but simultaneously constituted by discrete functional elements (namely spikes). Although neurons integrate input (resembling analog integrators), the output is often simply an all-or-none spike, depending on a threshold. The functionally significant variables in neuroscience—firing rates and spike timing—are peculiarly neural phenomena and do not correspond to the continuous variables manipulated by strict analog computers. Because neural signals are not precisely analogous to the continuous variables in strict analog computers, new mathematical theory specific to neural processes had to be invented—the mathematics of analog computers is not directly applicable.

These claims also support the broader philosophical endeavor of considering computation beyond the binary digital/analog framework. The possibility of non-digital and non-analog systems suggests that if there are representational types that are neither analog nor digital, one can formulate principled accounts of new kinds of computation. The debate over whether neural formats are analog, digital, or generic reveals that the notion of 'format type' is itself complex and multilayered. This naturally raises a meta-theoretical question: at what level of abstraction should we analyze formats? The same physical system can be described at multiple levels—from concrete physics to abstract computational specifications (and accordingly one arrives at differently constrained ‘wiggle room’ - that is,

degrees of freedom). How do these levels relate, and how does format operate across them?

6. Levels of Abstraction: From Interface to Implementation

6.1 Hoffman's Theory

Finally, we must consider the relationship between format and the level of abstraction at which a system operates. This brings us to Donald Hoffman's Interface Theory of Perception. Hoffman argues that our perceptual systems don't represent objective reality veridically, but instead function like a user interface—similar to how desktop icons on a computer don't resemble the underlying computational processes they represent. On this view, perception evolved not to show us the world as it truly is, but to provide a useful, fitness-relevant interface for guiding behavior.

Consider the desktop metaphor in modern operating systems. When you see a blue folder icon labeled "Documents" on your screen, this bears no resemblance whatsoever to the actual computational processes, voltage patterns, and magnetic states that constitute the physical implementation of that file system. The icon's color, shape, and position are arbitrary conventions. Yet this interface is extraordinarily useful precisely because it doesn't show you the underlying reality. You can organize files, move them between folders, and delete unwanted documents without understanding anything about memory allocation, file system architecture, or the physics of solid-state drives. The interface abstracts away irrelevant complexity and presents information at exactly the level of abstraction needed for your behavioral goals.

Hoffman argues that perception works the same way. Our visual experience of colors, shapes, and spatial arrangements doesn't reveal the true structure of external reality any more than desktop icons reveal the structure of computer memory. Instead, these perceptual experiences constitute a fitness interface—a species-specific, task-oriented representation shaped by evolutionary pressures to encode information about fitness-relevant affordances rather than objective physical properties. A snake doesn't need to represent the electromagnetic spectrum accurately; it needs to represent thermal signatures that indicate prey. A bee doesn't need a veridical map of three-dimensional space; it needs to represent the location of flowers relative to its hive.

6.2 Martinez's Critique: The Necessity of Integration

However, this theory has met with significant resistance. Manolo Martinez (2019b) offers a distinct, game-theoretic refutation of Hoffman's premises. Martinez accepts Hoffman's mathematical demonstration that "interface strategies" (which hide the truth to maximize utility) dominate in simple simulations, but he argues that these models rely on a fatal oversimplification: they assume agents are wholly "cue-driven"

In Hoffman's standard model, a sender (perception) observes a single cue and sends a signal to a receiver (action), which produces a payoff. In this linear, "narrow-band" context, it is indeed mathematically efficient to collapse distinct world states into a single "utility" signal, discarding objective structure. However, Martinez points out that biological reality is rarely this simple; organisms typically inhabit "informationally translucent" environments where utility cannot be determined by a single cue alone. And when the informational picture becomes more complex and

evidence channels mushroom, the efficiency-maximizing strategies that cut corners on veridicality fall apart[41].

To demonstrate this, Martinez models a "broad-band" agent, such as a bark beetle, which must integrate two distinct sources of information to survive: olfactory cues (chemical signatures) and visual cues (bark color). In this scenario, the relationship between the world and fitness is one-to-many. A specific chemical signature might be high-fitness in a "dark bark" context but low-fitness in a "light bark" context.

This necessity for integration exposes the fragility of Hoffman's Interface Theory. If the perceptual system employs a Hoffman-style "interface strategy"—collapsing world-states based on current utility—it destroys the structural information required to integrate the signal with other sources. If the beetle's olfactory system simply signals "good" or "bad" based on an average guess, it renders the visual system's input useless, because the "goodness" of the smell was contingent on the visual context to begin with.

Martinez demonstrates mathematically that in these synergistic scenarios, a "realist strategy"—one that preserves the objective structure of the cue rather than its immediate utility—delivers significantly higher fitness (an expected utility of 0.66 compared to the interface strategy's 0.44). To maximize fitness, the organism must "tell it like it is," preserving the causal structure of reality so that distinct sources of information can be combined downstream to guide precise action. Thus, far from driving perception away from truth, the need to integrate multiple sources of information drives representations toward veridicality.

---

[41] Another way to attack the Hoffmanian approach was suggested by Miłkowski (personal communication) on the basis of the Good Regulator Theorem (Conant & Ashby, 1970), which - to recapitulate very coarsely for our purposes here - proves that a control mechanism has to mirror the causal structure of the environment to work effectively. As perception is functionally an action-driving system, on any sufficiently bigger scale, it has to represent veridically. In other words, the 'gimmicky' strategies cannot match the payout of truth-tracking.

This theoretical vindication of the 'realist strategy' establishes a critical architectural constraint for the Neo-IPP: the cognitive system must possess a mechanism capable of integrating distinct information streams (e.g., context and content) without merging them destructively. If the system cannot distinguish the 'visual evidence' from the 'olfactory context' before integration, it cannot track the causal structure of the environment. As we will demonstrate in Chapter 4 (Section 6), the mammalian neocortex has evolved precisely such a mechanism already on a very low level of implementation: the 'two-point processor' architecture of the pyramidal neuron, which physically separates driving inputs from modulating contexts to solve the integration problem

Speaking of levels, Martinez's game-theoretic analysis demonstrates *why* the system must preserve the structural features of the signal (veridicality) to maximize fitness. However, to understand *how* these structural requirements are implemented across the hierarchy of the cognitive system—from the abstract algorithm of integration to the wetware that supports it—we need a framework for mapping these constraints. This necessitates a move from the flat analysis of utility to a layered analysis of validities.

6.3 When Formats Clash

That connects powerfully to Giuseppe Primiero's (2020) work on differentiating layered levels of abstraction in computational systems[42]. For Primiero, the formal level concerns what is being computed—the abstract mathematical specification of

[42] To be precise, Primiero is not the first philosopher to cash out analysis of computational systems in the terms of 'levels of abstraction.' Of course, there's Marr as the originator of this type of analysis, but also Luciano Floridi advocated for what he calls 'levellism' in a series of papers. This term seems to appear first in a 2004 paper co-authored with JW Sanders (which remains unpublished, to my knowledge). But then two highly influential papers, Floridi (2008) and Floridi (2010), introduce this avenue for understanding computational pipelines, and both predate Primiero's book.

a computational process. The physical level concerns how that computation is realized in actual hardware or wetware, with all the attendant constraints of real implementation. The experimental level concerns how computational systems are used as epistemic tools for understanding phenomena beyond themselves. As he explains in the preface to his 2020 book, this method is "recurrent throughout the present investigation, and it is particularly important in our definitions of computational validity, where appropriate variables of interest are identified: language and encoding for the formal version; architecture and behaviour for the physical version; realization of the formal structure in the physical implementation, and relation with models for the experimental version" (Primiero, 2020: x[43]).

Primiero's levels of abstraction framework provides a principled way to understand why representational format matters beyond mere information content. Format plausibly might determine factors such as which levels of abstraction are computationally accessible, what counts as validity at each level of analysis, why redundant formats might be systematically necessary, how information is preserved or transformed across abstraction levels, and the relationship between formal specification and physical implementation. This connects the philosophical debates about mental representation (propositional vs. imagistic, Cartesian vs. polar) to concrete questions about computational architecture and cognitive implementation. It suggests that the search for a single "correct" format may be misguided—instead, cognitive systems may need multiple formats operating at different levels of abstraction to achieve validity across the mathematical, physical, and functional dimensions of computation.

The necessity of distinguishing formats becomes most acute when we consider how different representational systems within the same agent interact. We have just argued for three claims in the previous chapter: some motor representations

[43] Sic! This is not a placeholder, but p. 10 of the Preface.

represent outcomes; there are actions whose directedness to an outcome is grounded in motor representation; and motor representation differs from intention regarding their representational format. A consequence of these claims is that a single purposive action may involve representations of the outcomes to which it is directed in at least two different representational formats: motor and propositional. This contributes to a problem Butterfill and Sinigaglia (2014) call the interface problem following Hoffmann's ideas (2016; more on that in 4.3).

Imagine that you are strapped to a spinning wheel facing near certain death as it plunges you into freezing water. To your right you can see a lever, and to your left there is a button. In deciding that pulling the lever offers you a better chance of survival than pushing the button, you form an intention to pull the lever, hoping that this will stop the wheel. If things go well, this intention will result in your reaching for, grasping, and pulling the lever. These actions may be directed to specific outcomes by virtue of motor representations that guide their execution. It shouldn't be an accident that, in your situation, you both intend to pull a lever and you end up with motor representations of reaching for, grasping, and pulling that very lever. If this match between outcomes variously specified by intentions and by motor representations is not to be accidental, what could explain it?

The interface problem is the problem of answering this question, of explaining how intentions and motor representations, with their distinct representational formats, are related in such a way that, in at least some cases, the outcomes they specify non-accidentally match. As Butterfill and Sinigaglia argue, postulating a process of translation is problematic because nothing at all is known about this hypothetical translation between intention and motor representation, nor about how it might be achieved. If the formats are truly distinct—for instance, if intentions are propositional and conceptually structured, while motor representations are

kinematic and non-conceptual—then a simple translation algorithm would require a "common code" that seemingly defeats the purpose of having distinct formats.

Instead, they propose a solution that hinges on the existence of demonstrative concepts (or other constituents) that refer to actions by deferring to motor representations. Just as the instruction "Follow this route!" succeeds in referring to a route by deferring to a cartographic representation (a map), so too can an intention refer to a motor representation. The intention does not need to encode the kinematic details of the movement; it needs only to point to the motor representation that does. This deference mechanism allows for coordination between systems without requiring a full translation of content, effectively bridging the format gap by functionally linking the high-level goal to the low-level execution.

6.4 Practical Concepts

This line of reasoning is extended by Carlotta Pavese (2021), who argues for the necessity of practical concepts to solve the interface challenge. Pavese distinguishes between practical reasoning (reasoning about what to do) and productive reasoning (reasoning about how to produce a task). While standard practical reasoning might conclude with an intention like "I will drink the coffee," productive reasoning must conclude with a motor command capable of guiding the specific movements required to lift the cup.

Pavese argues that for an agent to successfully perform a skilled action, they must possess productive intentions—executable plans that represent the task in terms of commands executable by the subject's motor system. Crucially, these intentions must be composed of practical concepts: representational units that are distinct from standard theoretical concepts because they are constitutively linked to the motor capacities of the agent. A practical concept of "grasping" is not just a descriptive category; it is a rule for generating the motor activity of grasping. By

positing practical concepts as the constituents of productive intentions, Pavese provides a mechanism by which the propositional format of thought can interface with the motoric format of action. The practical concept serves as the Janus-faced entity that looks upward to the intention and downward to the motor schema, ensuring that the "clash" of formats is resolved into a coherent stream of behavior.

This interface challenge underscores why format is not just a "gloss" or a metaphor. The difference between a propositional format (intention) and a motoric format (execution) creates a real causal gap that the cognitive architecture must bridge. The solution can be achieved, for instance, either via deference (see Butterfill & Sinigaglia) or practical concepts (as here with Pavese), but regardless of the choice, it minimally requires acknowledging that the mind traffics in distinct, incompatible formats that must be somehow coordinated.

This intuition of strict boundaries and worries about the commensurability (or at least communicability) between processing operating on differently formatted types of information has long fueled the philosophical treatments of this notion, which underscored the intuitive differences between icons and sentences or auditory and visual cues. This runs somewhat against (or at least independently of) some empirical work on multimodal integration, motor control, and “conceptual” spaces in AI research (Buckner, forthcoming), which simply conceptualized input from different modalities as types of data to be accommodated within high-dimensional spaces (and integrated there).

6.5 Pluralism and Redundancy

However, while such behavioral interventions provide the gold standard for investigating format in simpler organisms, humans present greater complexity. Unlike desert ants with their narrow behavioral repertoire and easily disrupted navigation, humans exhibit flexible spatial behavior across virtually any terrain and

can effortlessly translate information across scales and modalities. This flexibility likely requires multiple coexisting navigational systems rather than a single, specialized format. Evidence increasingly suggests that humans maintain both Euclidean/Cartesian and graph-like/polar representations simultaneously, calling upon different formats depending on task demands.

This leads us to the broader principle of architectural redundancy. Rather than constantly translating information from one privileged format into task-appropriate formats—a process that would be computationally expensive—the mind may simply maintain multiple formats in parallel. This idea is not new and was defended most famously by Paivio (1986) under the banner of Dual Coding Theory. Dual coding posits that any given piece of information can be formatted both visually and verbally, representing an early and influential recognition that the mind might maintain multiple representational formats simultaneously. On this view, the concept "chess," for instance, may have both a visual form—perhaps an image of a chessboard with its distinctive black and white squares and recognizable pieces—and a verbal form consisting of explicit, language-like knowledge about the organization of the board and where each piece belongs.

This dual representation serves multiple functional purposes. Most obviously, having information encoded in two different formats may reduce the likelihood of forgetting: if one trace degrades, the other may remain accessible. Furthermore, different formats may be better suited to different cognitive tasks. The visual representation might support rapid pattern recognition during gameplay, while the verbal representation might facilitate explaining strategies or teaching the game to others. The adaptive value of such redundancy becomes particularly clear when considering cases of selective impairment. If information exists in only a single format and that representational system is damaged, the information is lost entirely. But if the same information is stored redundantly across multiple formats,

damage to one system need not eliminate access to that knowledge. A patient who suffers damage to visual processing areas might lose the ability to visualize a chessboard but retain verbal knowledge of chess rules and strategies.

6.6 Neural Reuse

This demand for efficient formatting finds a biological parallel in the phenomenon of neural reuse. Traditionally, the philosophical argument against hard structuralism was multiple realizability (Putnam, 1960; 1967)—the idea that the same function can be implemented using entirely different physical systems. More neurally inclined philosophers that accepted multiple realizability felt vindicated in their positions by the discovery of neural plasticity in the brain. If there is actual evidence that the same brain structures can be reemployed to perform different functions during the life cycle of the entire organism, that provides strong support for the idea that structural organization is secondary to function.

Recently, however, a new understanding of neural recyclability (redeployment) has emerged. Called 'neural reuse' by Michael Anderson (2007, 2010), it was hailed as a new organizational principle of cortical parcellation. The main difference from the old-fashioned neural plasticity lies in the very permanence of the change. According to neural reuse, circuits can continue to acquire new uses after an initial or original function is established; the acquisition of new uses need not involve unusual circumstances such as injury or loss of established function. Reuse is treated as a run-of-the-mill organizational principle in which healthy, normally functioning structures routinely take part.

The given circuit participates in a few separate bigger structures (such as the resting state network), just during different timescales. In other words, there is overlap of different networks in the area of one circuit. This mechanism suggests that the brain is highly efficient at repurposing representational vehicles. Rather

than building new hardware for every new task, the system reformats the inputs and outputs of existing circuits to serve new functional roles. This flexibility is a flexibility of *format*—the same physical substrate can support different informational codes depending on the network embedding and temporal context. This reuse mechanism provides the physical substrate for the redundancy we observed in spatial cognition: the same neural resources can be formatted to support different coordinate systems depending on the task context.

These empirical findings—that humans maintain multiple coordinate systems simultaneously, that format affects systematic error patterns, and that redundancy serves crucial functional purposes rather than representing waste—raise a more profound question about the nature of neural formats themselves. We have established that formats matter empirically and that multiple formats can coexist. But what type of representational format do neural systems actually employ? This question has generated sustained philosophical debate, with three major positions emerging. Some argue that neural representations are fundamentally analog, exploiting continuous physical magnitudes. Others contend they are essentially digital or symbolic, implementing discrete computational operations. Still others propose that neural computation is sui generis—a unique type irreducible to either traditional category. Understanding this debate is essential because the answer has profound implications for how we understand neural information processing, what computational operations are available to biological systems, and how Marr's levels of analysis apply to neuroscience.

7. Implementation & Granularity

Speaking of Marrian lenses of the computational, algorithmic, and implementational, let’s start from the bottom and talk about the instantiation -

what are neural vehicles? As one suspects, there are competing factions even here. And again, the mainstream view (we can call it the 'classical neuron doctrine') spearheaded by Hubel and Wiesel's view on selectivity and tuning curves, supplemented by the HH (Hodgkin-Huxley) model of the axon and belief in spike-train coding (and, well, Shannonian information presiding over all of them), seems to be staggering, too. The vultures are circling.

Since there is no longer a monopoly on the market of neural implementation, now I want to review *three* main families of theories postulating the right level of explanation for neural information processing systems: the neuronal, the populational, and the sub-neuronal. This might seem like a futile endeavor in a philosophical text, as it cannot reach either the depth or sophistication of neuroscientific literature, while also seeming relatively unimportant for discussions on information processing at a glance. The worry here would be along broadly functionalist lines - that implementations hardly matter for trying to find and police philosophical divisions and cognitive modules.

But this is wrong. As we mentioned, informational explanations face the main challenge in aligning f-information with a-information. To do so, they must do whatever is possible in their power to demonstrate that their recording, decoding protocols, or any other procedures aimed at arriving at actual informational substrate, cannot be unfeasible for actual internal consumers that use them, for which it is explicit, exploitable, and functionally relevant.

To understand this challenge, we might notice—perhaps - perhaps counterintuitively, that researchers might use too powerful a tool in this case. This might seem weird, as it would amount to an astronomer lamenting a telescope too sharp or an oceanographer a sonar too precise. And of course, for certain pragmatic purposes, that might be unwieldy, costly, or not worth the effort. But in

terms of neural decoding, there is an in-principle case for that. And interestingly, the practitioners agree. It is for this reason why virtually nobody claims that using non-linear classifiers to parse through neural data should count as an agreeable way to decode actual dimensions of difference (for the system). These systems, while powerful, at detecting even the slightest tendencies and patterns in the data, are certainly way over the possibilities of any biologically realized linear classifier that might hide within a mammalian skull.

But the problem goes deeper than mere computational power. The problem with these types of models is that they are what we call 'informationally greedy.' Put intuitively, they are shameless devourers of any tiny pattern within the dataset and do not shy away from arriving at the decision. This feature could be extremely helpful if what we are after are blunt results - successful classification, categorization tasks, and predictions. But as explanations and models of behavior, these almost certainly fail, as they far outcompete the possibilities of biological systems (in their very limited tasks). In a way, this is a problem analogous to the need for psychologically plausible models of how downstream processing functionally exploits that information for guiding behavior that we discussed earlier. Thus, we might see why the notion of the biological plausibility of the information processing pipelines one postulates within the neural environment is paramount. We can now turn to analyzing the main leading theories of neural computation.

7.1 The Neuron Doctrine

The neuron doctrine represents the classical and widely influential idea that the individual neuron is the basic functional, signaling, and computational unit of the nervous system (Rolls, 2021). Historically, this view gained prominence following the pioneering work of McCulloch and Pitts (1943), who connected the all-or-none

law of nervous activity to computation (Piccinini, 2021). They formalized the neuron as an input-output device, suggesting that the activity of any neuron could be represented as a simple proposition because its activity follows the "all-or-none" principle (McCulloch & Pitts, 1943). Neurons are widely assumed to be the vehicles of information processing, primarily through the production of spikes (action potentials) organized into sequences called spike trains (Piccinini, 2021; Rolls & Treves, 2011).

The classical mechanism assumes that a neuron operates as an integrate-and-fire point processor, summing all its excitatory and inhibitory synaptic inputs to compute a single net activation level, and transmitting information if that sum exceeds a threshold (Phillips, 2023). This simplification persists because neural nets composed of such simple units are, in principle, capable of computing anything computable (McCulloch & Pitts, 1943). The neuron doctrine fosters an abstraction away from many complexities of cell biology, focusing instead on whether the logic gate functions are fulfilled (Piccinini, 2021). Despite its utility, this traditional view has faced criticism, particularly the radical version, which asserts that explanations must bottom out exclusively in neurons or molecules (Gold & Stoljar, 1999). However, effective explanations in modern neuroscience often span multiple levels of mechanisms (Craver, 2007). However, the neuron model has encountered serious challengers from both above and below. First, let’s look above.

7.2 Level Up: The Population Doctrine (Hopfieldianism)

The Population Doctrine, frequently dubbed the Hopfieldian approach (Barack & Krakauer, 2021; Ebitz & Hayden, 2021), shifts the explanatory spotlight away from individual nerve cells and towards patterns of activity distributed across entire neural populations (Kinsey, 2023). This doctrine insists that the primary unit of functional analysis is the population of neurons, rather than its distinct constituent

cells (Ebitz & Hayden, 2021). The characterization of this approach focuses heavily on distributed patterns of system activity and the development of models capable of capturing the global dynamics of the system (Willett et al., 2020)[44].

Under this view, representations are conceived of as trajectories within a high-dimensional state space, where the firing rate of each neuron defines an axis (Shea, 2018). In this context, activity is sometimes observed to be confined to a low-dimensional manifold spanned by a few independent patterns known as "neural modes," which capture a significant fraction of the population covariance (Richards et al., 2019). The application of this doctrine often utilizes advanced methods like dimensionality reduction techniques (e.g., PCA) to derive mathematical summaries of how population activity fluctuates with task parameters. Proponents of this view argue that these analyses demonstrate that neural systems represent information in a distributed manner, rather than a localized one, which challenges certain algorithmic models of brain function (Burnston: 2021a; 2022).

These types of views were emboldened by the groundbreaking study by Mante and colleagues (2013). The study seemed to demonstrate that in the face of modern more thorough experimental protocols, the received, venerable picture of stable selective receptibility. Historically, it relied heavily on the concept of the receptive field (RF; Marr, 1982), tied invariably and hyper-selectively to a particular type of stimulus—a view that no longer seems tenable, even for single neurons. The RF concept defines the specific portion of the sensory world that can influence a neuron's activity (Rieke et al., 1997). This idea was established by studies showing neurons function as detectors for specific configurations, such as center-surround organization in the

[44] This study is a landmark case, because the authors show that the "hand knob" area—previously thought to only control the hand—contains a global, distributed code that represents movements of all four limbs (legs, arms, face) through a "compositional" neural encoding. Thus, it provides evidence for where neural activity is viewed as a trajectory through a high-dimensional state space rather than as individual firing rates.

retina (see: Barlow, 1953; Kuffler, 1953 for classic formulations) and specialized responses in the visual cortex (Hubel & Wiesel, 1962). These discoveries suggested that questions of psychological interest could be illuminated, and perhaps even explained, by observing the responses of single cells. If a cell acts as a reliable "bug detector" (Lettvin et al., 1959) or an "oriented edge detector" (op. cit.), it is seen as the dedicated physical vehicle carrying that univocal content (Burnston, 2020).

In detail, the central purpose of the Mante study was to examine PFC activity in monkeys trained to flexibly select and integrate noisy sensory inputs (motion and color) based on a specific behavioral context, ultimately leading to a choice (Mante et al., 2013; Burnston, 2020). Their results pointed not to stable classical encoding properties, but towards distributed and complex encodings. The activity of individual prefrontal neurons was remarkably complex, generating responses that defied deep understanding if analyzed in isolation. In turn, the researchers found that information pertaining to motion, color, context, and choice were all distinguishable at the population level of activation. Also, the neurons in the population exhibited widespread ‘mixed selectivity’ (Rigotti et al.,2013). This means that rather than being specialized for a single experimental variable, individual cells were selective for multiple task parameters simultaneously. They interpreted their result as that the observed complexity and the functional roles of single neurons were best understood only when viewed through the framework of a dynamical process unfolding at the level of the population. The mechanism responsible for both the selection of relevant input and the integration of that input was determined to be two aspects of a single dynamical process unfolding within the same prefrontal circuits (Mante et al., 2013). This suggested - according to them - a previously unknown, holistic mechanism for context-dependent computations.

7.3 Sub-neuronal Accounts

However, the challenge to the explanatory supremacy of the single neuron is bidirectional: just as some theorists look upward to the population, others look downward into the cellular machinery itself. Sub-neuronal accounts challenge the simplicity of the point neuron model, emphasizing that significant information processing occurs within the detailed, complex structure of the neuron itself, or even at the molecular level (Koch, 1999)[45]. This perspective shifts the focus beyond the cell body to understand how the physical features of neural representations are manipulated by physical neural processes (Maley, 2024). A major component of this approach is 'dendritic computation,' which reveals that neurons are far more than simple summing nodes (Poirazi & Papoutsi, 2020). For example, conductance-based models, which describe voltage evolution based on ion channel densities and morphology, show that the Hodgkin-Huxley model, while simple, performs a computation of "surprising richness" that is not well described by a single threshold (Agüera y Arcas & Fairhall, 2003; Fairhall, 2014).

Crucially, *Cooperative Neuron* models show that a common class of neocortical pyramidal cells operates with two functionally distinct sets of input (Phillips, 2023). 1. One set provides information about what the neuron usually transmits (receptive fields/basal dendrites). 2. The second, highly diverse set provides context (contextual fields) typically arriving via the apical dendrites in layer 1 (Schuman et al., 2021). These neurons utilize this context sensitivity to selectively amplify the

---

[45] Interestingly enough, a fair share of machinery in the neurons, including these directly responsible for spike generation (recovering membrane excitabilities, gating potentials, etc.), are realized by non-electrical, purely chemical processes. In a way, the dynamics of cell activity are the constant flux between electrical and chemical systems balancing each other. These processes ARE computationally modeled routinely (as even a semester NEURON course graduate knows; I can personally vouch for that). However, it is not customary to explain them as computational in the literature.

transmission of information about their basal input when coincident excitation occurs at both sites (ibidem). This cooperative context-sensitivity increases the coherence and usefulness of the system's overall activity, challenging the "dendritic democracy" assumption in the neocortex (Magee & Cook, 2000). Finally, sub-neuronal accounts also advocate for analyzing mechanistic processes at finer scales, noting that some philosophers recommend expanding the list of functional units beyond neurons to explicitly include synapses and glia[46] (Cao, 2014).

8. Double Down: The Critique of the Point Neuron

While sub-neuronal accounts correctly identify the computational richness available at the microscale, we still require an architectural principle that organizes this biophysical complexity into a functional processing unit. The dominant "point neuron" doctrine used in systems neuroscience fails to capture this organization, implying a dendritic democracy where all synaptic inputs—regardless of location—compete on equal footing to determine the cell's output.

Phillips (2023) critiques this simplifying assumption. Rather than abstracting away the anatomical complexity, he doubles down on it, revealing that the pyramidal neuron is not a single point but a two-point integrator. By respecting the strict biophysical asymmetry between driving inputs (content) and modulating inputs (context), this model solves the computational bottleneck of the point neuron. The crux of this architecture is apical amplification—a mechanism that allows the

[46] Currently in neuroscience, there is a fascinating research program bubbling that aims to investigate astrocytes (which were traditionally thought to be supplementary metabolic cells, not implicated in signal transmittance) as information processing systems.

neuron to boost relevant signals without corrupting their representational content. We will defer the rigorous mechanistic treatment of this solution to Chapter 4 (Sections 6.1 and 6.2). We do this for a specific theoretical reason: apical amplification is not merely an alternative format definition; it is the specific hardware requirement for naturalizing *content*. With this forward pointer in place, we can now conclude our analysis of format by summarizing how these constraints shape the determination of content.

9. Conclusions - Philosophy of Format Redux

This account of how format constrains content determination through accessibility structures brings us back to the question that opened this chapter. Why do we need the notion of format at all? Could it be eliminated in favor of purely contentful or purely implementational explanations?

This distinction became foundational for understanding how representational format matters independently of representational content, and it remains central to debates about mental representation, external representations, and cognitive efficiency. While one might dismiss appeals to format as reflecting a lazy reliance on computer metaphors—treating mental representations as analogous to differently formatted digital files—this concern underestimates the concept's genuine explanatory utility. Format distinctions are, in fact, employed widely throughout the cognitive science literature and perform important explanatory work in contemporary research. As David C. Cox (2014: 189) observes, decoding research on the visual system implicitly recognizes that the fundamental problem of vision is not primarily one of information content, but rather one of format. We know that the activity of retinal ganglion cells contains all the information that the

visual system can use. The computational task is reformatting it for downstream use.

Ultimately, formats provide substantial explanatory purchase precisely because they capture how information structure constrains and enables cognitive processing. The format in which information is encoded determines what computational operations can be efficiently performed on it, what inferences can be drawn from it, and how it can interface with other representational systems. This explanatory role cannot be easily eliminated without sacrificing our ability to account for key aspects of cognitive architecture and neural computation.

Format, we can now see, is a constraint on content determination, on computational efficiency, on neural implementation, and on cognitive architecture. Understanding cognition requires understanding format not as a peripheral concern or convenient metaphor, but as a constitutive feature of how information is structured for use. The brain's formats are its solutions to the problem of making information explicit, accessible, and exploitable—and those solutions shape what the brain can know and do.

Chapter 4: Content

1. The Structural Problems

1.1 Directionality of Information and the Arrow of Time

We have already mentioned Fodor's points about the importance of asymmetric relations for fixing the semantic properties of cognitive systems. However, current contemporary cognitive science approaches directionality quite differently. ‘Directionality’ in terms of information flow typically refers to the direction in which signals propagate through systems (feedforward vs. feedback), the direction of causal influence in neural networks, or the temporal ordering of information processing. We are interested in how information is processed downstream and in cognitive pipelines in general. This is usually operationalized empirically through techniques like Granger causality, transfer entropy (Granger, 1969), directed transfer function (DTF; Kamiński & Blinowska, 1991) or structural equation modeling (Bollen, 1989).

That is why the most canonical explanations of major cognitive subsystems rely on identifying 'pathways' that we can think of intuitively as sort of informational highways (or data buses, in more information engineering terms). Customarily, they are cashed out in Shannonian, directed terms. But mutual information is another measure used heavily in interpreting the relationships between two sources of neural data.

The formula for mutual information (MI) between two random variables X and Y is:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

where:

- $H(X)$ is the entropy of X
- $H(Y)$ is the entropy of Y
- $H(X, Y)$ is the joint entropy of X and Y

This can also be written as:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

where $H(X|Y)$ is the conditional entropy (uncertainty about X given Y).

Mutual information (MI) serves as a valuable, non-linear measure of statistical dependency or association between variables, quantifying the amount of shared information, often measured in bits. However, standard MI is an information-theoretic measure of functional connectivity that is inherently descriptive and non-directional, meaning it reflects statistical dependencies but does not directly infer underlying causal influence or coupling. This non-directional nature significantly limits its explanatory power in contexts, such as neuroscience, that demand a mechanistic understanding of system interactions.

MI is symmetrical and only assesses the degree of relationship, failing to account for how correlations may arise from common inputs rather than direct coupling between two components. Therefore, demonstrating a change in functional connectivity or MI between two areas does not necessarily imply a change in the underlying coupling strength, as it can be confounded by differences in signal or noise. When the scientific goal shifts from merely describing associations to characterizing directed interactions or effective connectivity, non-directional measures like MI are insufficient. Effective connectivity requires explicit causal modeling, such as Dynamic Causal Modeling (DCM; Friston, Harrison & Penny,

2003), which tests hypotheses concerning coupling architectures using generative models.

Even measures explicitly designed to assess directionality based on information flow, like Granger Causality (GC) and Transfer Entropy (TE)—which are closely related to MI[47]—face conceptual problems in explanatory power. GC, for instance, is often indifferent to the dynamics of the receiving node ("receiver independence" - pointing directly to the receiver relativity problems we discussed in Chapter 1) and the amplitude of effects, reflecting only a combination of transmitter and channel dynamics. Consequently, causality analyses like GC can be difficult to interpret and may not align with the mechanistic understanding of neurophysiological processes that seek to explain observed effects at a particular site.

The explanatory power of mutual information is significantly enhanced only when it is rigorously restricted and contextualized by directional and functional constraints, often seen in the theory of maximal mutual information (maxMI). The maxMI theory aims to generate content attributions suited for mechanistic explanations by invoking non-etiological functions (C-functions; Heemskerk, 2025) to restrict the scope of MI analysis. This approach isolates the element of the external environment (modeled as an external random variable, eRV) with which an internal representation (modeled as an iRV) shares maximal mutual information, specifically within the functional profile defined by the system's capacity to process the information. This restriction to usable information (information decodable by

---

[47] To be exact, mutual information I(X;Y) measures how much knowing X reduces uncertainty about Y, and it's symmetric: I(X;Y) = I(Y;X). Transfer entropy adds directionality and time by incorporating time lags and conditioning, making it asymmetric—it answers, "Does X's past help predict Y's future beyond what Y's own past tells us?" Granger causality is essentially the linear special case of this relationship: for Gaussian processes, Granger causality and transfer entropy are mathematically equivalent, with Granger using prediction error reduction in autoregressive models while transfer entropy generalizes this to nonlinear relationships using information theory. The key advantage of transfer entropy is that it captures nonlinear causal relationships that Granger causality (which assumes linearity) would miss (Barnett et al.,2009; Hlaváčková-Schindler et al.,2011).

downstream components) is crucial because it ensures that the resulting content is available to the system itself, meaning its values can undergo processing by the causal mechanisms. By integrating the directional constraint of the C-function with MI maximization, maxMI ensures that the content isolated is explanatorily salient for the mechanistic projects of cognitive science, linking the external item through an identifiable information processing chain to the eventual internal decoder.

The temporal aspect is crucial when moving beyond non-directional measures like standard mutual information (MI) to analyze the directionality of information transfer. However, Barwise and Seligman raise a critical caution against symmetrical measures such as MI. I will call it the 'Arrow of Time' argument:

'Causal relations often underpin informational relations, but they are not the same. One important difference is that the direction of information flow is not necessarily aligned with the direction of causation. Present events can carry information about conditions in the remote past and about their future consequences, but there are strong arguments against the possibility of a present event causing a past event' (Barwise & Seligman, 1997: 18).

We can see here that the flexibility of information-bearing relations might actually be detrimental to their ability to carry explanatory loads. Archeologists routinely infer the properties of ancient cities based on properties of, say, states of current-day ruins, texts, or even geographic features (as in the case of Troy excavations). Present tokens are informative about the tokens buried deep in the past. But that clearly violates the usual order of things assumed by the causal theories—they only work forward, not backward. For this reason, we can assume that unconstrained measures of information can be helpful in terms of assessing connectivity and even potential semantic relations (what type of information is

being sent and exploited downstream), but they cannot serve as a definitive basis for this type of content ascription.

1.2 The Indeterminacy Problem: Neurorabbit

A classic philosophical parable of a field linguist trying to understand the linguistic practices of 'rabbit country' may seem like quite a stretch from philosophy of cognitive science, or especially – neuroscience. Since the radical translation problem is concerned with the personal-level, linguistic communication and ascertaining intentions of the speakers with the effect on listeners; this problem does not seem to fit inside the head of a single person. However, as pointed out by Rosa Cao (2017), when we tackle the problem of content ascriptions in the mind and more generally—favoring one explanatory theory above the others—we might find ourselves back in the rabbit hole.

To recall the problem: Quinean indeterminacy of translation originates from Willard Van Orman Quine's arguments from *Word and Objec*t (1960) regarding the radical translation of linguistic expressions. Sometimes called the 'inscrutability of translation.' The principle suggests that multiple rival interpretations of any linguistic signal's meaning or content remain equally valid because they successfully account for the observable responses of the speakers. Whenever a newcomer (such as Quine's linguist-explorer) finds himself among the speakers of the unknown language and tries to reconstruct the meanings from scratch. No amount of observations can make him escape the arbitrariness of his assumptions. After all, even when the natives point at the rabbit yelling 'Gavagai,' he cannot indisputably assess whether what they meant was really "Lo, rabbit" or rather "an undetached-part-of-the-rabbit"[48].

---

[48] It is worth noting that the inevitability of this inscrutability has been challenged on logical grounds. Zabłudowski (1982) offered a powerful critique, arguing that Quine's thesis is internally contradictory. If one strictly accepts Quine's behaviorist premise—that meaning is exhausted by

Quine's idea, when neuralized (meaning 'applied to neural systems') posits that the assignment of determinate semantic content to internal brain signals cannot be uniquely fixed by empirical neural and behavioral data alone. Cao argues that neuroscientists attempting to decipher the neural code find themselves in a position analogous to Quine's field linguist: constrained only by the stimuli that evoke the signal and the response the signal elicits. Consequently, there is no way to decide between competing claims about what information content or "meaning" that the neural signal carries, resulting in inevitable[49] inscrutability of translation, i.e. indeterminacy.

She utilizes this argument to critique ambitious semantic claims made by contemporary computational theories, particularly predictive processing models. By doing so, Cao highlights the informational equivalence between models with fundamentally different content assignments. For instance, traditional models might hold that bottom-up neural signals carry 'sensory information about their distal causes,' while predictive models insist these same signals convey 'error

---

'stimulus meaning'—then translation is fully determined. Indeterminacy can only arise if one implicitly accepts that meaning transcends stimuli. Thus, Quine does not prove indeterminacy; he merely demonstrates that his own concept of stimulus meaning is inadequate. Furthermore, Zabłudowski demonstrated that even logical connectives cannot be uniquely decoded from behavior alone—a finding that predates and anticipates contemporary arguments by Papayannopoulos, Shagrir and Fresco (2022) regarding the indeterminacy of computational states also in the context of neuroscience.

But instead of recapitulating their views, we can allow for a rare move and illustrate it with the anecdote found in Dunn (2008:9): 'I believe that the fundamental philosophical question regarding the representation of information is: how much "lies in the eye of the beholder?" This is meant metaphorically, of course—replace the word "eye" with "interpretation" or even "mind." This can benicely illustrated by an anecdote the late Australian philosopher Ian Hinkfus (sic!) once told me. He was working for IBM in the design of an early computer, and they ran out of NAND gatesin building their prototype. So they used NOR gates instead and just reinterpreted the output'. A final note here: Dunn was also probably mistaken there, as Copeland and Shagrir (forthcoming) demonstrate the engineer in question was not Hinkfuss, but a man named Ralph Slutz (I thank my supervisor for his archival alert in this affair).

[49] It was correctly suggested to me that there are high-level concepts (like "property of the Pope") that are not reducible to sensory stimuli, challenging the claim that neuroscientists are only constrained by stimuli. I acknowledge this, but I think the response for a *Neuroquinean* would be that the initial decoding of a neural signal is strictly bound by observable inputs.

signals' (information about the difference between predictions and actual input). Cao argues that, when quantified by standard statistical measures (like Shannon information), the concepts of sensory information and prediction error are 'informationally identical,' except that the latter is conceptually centered on predictions. Since the statistical information content is the same regardless of which label is applied, the neural data alone cannot force the choice between the predictive or the traditional interpretation; the resulting assignment of content remains arbitrary.

1.3 The Decoder's Dilemma (Brainjackers and Mindreaders)

This alignment problem is also known as the Decoder's Dilemma. As Brendan Ritchie and colleagues argue, we know that the activity of retinal ganglion cells contains all the information that the visual system can act upon. What's more, that nonlinearity and noise in neuronal processing can only decrease (and never increase) the absolute amount of information present. However, all this information present in the firing of retinal ganglion cells does not reside in a format that can be easily read out by a downstream neuron to guide action. Information that is decodable by Multivariate Pattern Analysis (MVPA) might still be implicit from the brain's perspective if it's not easily usable or multiply exploitable by downstream neural systems. The neuroscientist's sophisticated algorithm can extract information that is computationally expensive or impossible for the brain's own circuits to access.

So, as we saw from previous discussions, neural observers (receivers) galore. Obviously, the brain holds many parallel and interconnected systems that cater to different needs and realize separate functions. We conceded that implies the proliferation of contents and formats in play that have to be carefully pried apart to reliably eavesdrop on the intracranial chitchat. But most of this chatter stays in the

brain - with one significant exception already signaled by Ritchie and colleagues, the neuroscientists. But are these conceptual roadblocks really 'thwarting the brainjackers'[50]?

Granted, the situation where the receiver is not internal to the cognitive system but external, as in the case of neurorecording, is a peculiar one. For starters, it hardly fits the teleosemantic picture of the long-standing, evolutionarily driven chain of content ascription. There is no obvious benefit for the information transfer to be obtained between the brain and the brainjacker (even if he acts at a distance), at least from the organism's point of view. This link is usually very short, unstable, messy, and coarse-grained (from the normal neural signaling perspective). It also disregards common informational pathways employed.

This leads to a predictable problem. Is the information gained this way really the same one that the brain employs? This is obviously a huge assumption, but the conflation is omnipresent (de Wit et al.,2016[51]). As established in Chapter 1, we must distinguish between 'information-for-the-system' (f-information) and 'information-about-the-system' (a-information). We can also think of the format as 'information-in-the-system' as opposed to the one that is external to it. Thus, we could also think of it as an internal/external boundary.

Yet another potentially helpful way to think about these distinctions is to differentiate information as a measure from information as a substrate (a well-known debate in philosophy of information; see Floridi, 2013). It is especially tempting to think of neural information as the latter (and again, this is commonplace in empirical science), but this move - whilst not impossible -

---

[50] I am borrowing this amusing term from a title of a short blog note by Cavuto (2017).

[51] Buzsáki (2019) calls this difference the "brain from outside-in" (our external information) and the 'brain from inside-out' (one he advocates for; *internal* in our lingo)

requires cautious treading and fulfilling additional requirements. For instance, mutual information is even intuitively different from information transfer. Directionality is one of the reasons, but they may also vary in terms of rate and function, even if one direction of MI aligns with the actual implementational informational transfer pathway.

The goal of the informational program in cognitive neuroscience is then for f-information and a-information to align. In a sense, the entire program of informational explanations is geared towards making these two equivalent. Then, we can say that we - as external observers - utilize the same information (in terms of content and format) as its intended receivers. This is the end task of the neural decoding research program that intends to demonstrate that external a-information that we gain access to during our neurorecording procedures (supplemented by additional steps) is in fact the same as internal f-information (for a given task and subsystem).

But while this conceptual difference is of utmost importance for theoretical reasons, the question of aligning the two remains a steadily empirical question. What philosophy of science has to do is to provide a set of conditions or desiderata that should be cleared as necessary and sufficient auxiliary conditions (auxiliary to the recording itself).

So, how to advance from empty promises of brainjacking towards truly explanatory mindreading, as the IPP proponent ought to strive for?

2. The Normative Turn: Fixing Content

2.1 The Failure of Pure Correlation

Unluckily for a theorist interested in the explanatory power of informational theories, the 'interpretative glosses' identified by Cao[52]—whether we call a signal 'sensory evidence' or 'prediction error'—amount to little more than semantic decoration if they remain empirically indistinguishable. As long as these rival descriptions belong to the same statistical equivalence classes, they cannot do genuine explanatory work. To escape this impasse, we need more than just a different label; we need a mechanism that can generate divergent empirical predictions. This forces us to confront the limitations of purely correlational theories. If 'meaning' is nothing more than statistical correlation, we are left without a principled way to assign determinate content or to account for error.

This impasse is famously crystallized in Stevan Harnad's (1990) Symbol Grounding Problem. Harnad asked how the semantic interpretation of a formal symbol system can be intrinsic to the system, rather than parasitic on the meanings in the head of an external designer. If mental symbols are merely manipulated based on their shape (syntax), how do they ever touch the world? Pure computationalism, he warned, risks becoming a 'merry-go-round' of empty symbols pointing at other empty symbols (like a dictionary from Polish to Polish). Luckily, for the Neo-IPP, this is not a death knell but a design constraint: it demands that our theory of content must identify a mechanism that 'grounds' the symbol not in a definition, but in a causal transaction with the environment.

This may appear to be a tall task. For, *prima facie*, causal-informative theories (such as Dretske's early view) are hard-pressed to say how misrepresentation could

---

[52] This problem will be covered in more detail in Section 6.2.4 of Ch. IV.

emerge. After all, if the causal link is enough to fix the content, any (even aberrant or anomalous) circumstances that lead to the tokening of the sign would constitute its content. The mistaken or faulty tokening could not be interpreted as such; causal-informative theorists would have to accept the shift in the content.

Famously, this invited Fodor's (1984) critique of what he called 'Wisconsin-style semantics,' in which he demonstrated that the inability of these theories to disqualify errant causes leads to what he called the Disjunction Problem. In a famous example, when a representation HORSE is learned to mean 'horse' but is then triggered by a cow in a dimly lit environment (that leads to such misattribution), a causal theory is forced to shift the representation's content of HORSE to "horse OR dimly lit cow". Such a 'new' disjunction is then considered a 'new' correct representation of the target—now a disjunctive class, without the ability to disambiguate between this and the 'nomic' normal 'horse means horse' case.

2.2 Asymmetric Dependency

Fodor's response was to conjure the Asymmetric Dependency condition, which pries apart informational relations from representational ones on the basis that informational relations depend upon representational relations, but not vice versa. In other words, gimmicky instances of triggering the representations piggyback on ones brought about during normal conditions. Information about certain parts of 'horsiness' could be elicited by perceiving a horse-like-looking cow, but dimly lit cows do not participate in establishing the 'horse' representation. To understand that, it's useful to think of fakes, forgeries, or copycats. A counterfeit banknote produced by the master forger draws its effectiveness from the fact that such-and-such-looking pieces of printed paper are established to be legal tender. The relation is asymmetrical. Forged notes are similar to the original and vice versa

(they are mutually informative of their appearance, for example). However, only the original's veracity stipulates the relation, not the other way round—it's asymmetric.

3. Teleosemantics (The Backward-Looking Solution)

3.1 Biological Function

But apart from causal-informational theories, there is another strand of strategies for naturalizing content while preserving the possibility of representational errors - the teleofunctional theories (Millikan, 1984; Neander, 2017). According to teleosemanticists, representational relations are only those that were selected by a representation-producing mechanism via evolution or learning. As such, this mechanism, stipulated on a much longer timescale than in causal-informational theories, is granted by their function.

Crucially, the tokening of the representation is not defined by the distal object itself (the cow), but by the proximal stimulation it causes. A dimly lit cow does not signify 'dimly lit cow' simply because it caused the firing; rather, it triggers the representation 'HORSE' because it happens to satisfy the visual profile that the mechanism was selected to detect. The mechanism has the proper function of indicating horses. Therefore, the system that is triggered by the misleading visual profile of a dimly lit cow is not defining a new content; it is malfunctioning relative to its proper function. It is a case of a 'false positive' generated by a valid mechanism operating in invalid conditions.

Teleological theories differ from causal-informational ones in a few important dimensions - the timescale is much longer, but more importantly, the constitutive relation is not mediated by co-occurrence or causal mechanisms but proper

functions, which are more flexible. It loosens the requirement that a function must be instantiated successfully in every individual organism. Unlike causal theories, which require a history of successful tracking within the individual, teleosemantic functions are ascribed at the level of the population or lineage. A specific individual might never successfully track a horse, yet their neural mechanism still possesses the function of 'tracking horses' because that is what the mechanism was selected for in their ancestors. This allows for a normative standard (correct vs. incorrect) that exists independently of the individual's immediate causal history.

3.2 Liberal Representationalism

Hence, there is also a philosophical motivation for the strategy called by Artiga as 'liberal representationalism' (Artiga, 2016). Liberal representationalism wants to vindicate the broad usage of the term in science but also build it into philosophical groundwork. According to Artiga, it is the acceptance of the teleosemantic theory of content that not only opens up but actually logically leads to liberal representationalism. Let's use his syllogism:

1. TELEOSEMANTICS↠LR
2. TELEOSEMANTICS

   • Therefore, LR.

   By 'teleosemantics' he understands holding that 'R is a representation iff R stands within a sender-receiver structure, such that the biological function of the sender is to produce R when a certain state obtains and the biological function of the receiver is to produce an effect when R is produced' (Artiga, 2016:412) 5.

We can translate this reasoning for the selectively responsive case of visual cells. If these cells generate a specific pattern of spikes (as experimental data shows to be the case), when there is a such-and-such tilted bar in a given part of the visual field

and there are mechanisms further downstream (e.g., higher up in the layers of the visual cortex), whose function is to use this activity to accomplish other tasks (for instance, reconstructing shapes), then the activity of these cells counts as genuine representations according to the teleosemantic account. Artiga calls these kinds of very minimal states that seem to qualify as representation not only on practical terms but also on philosophical grounds of teleosemantic theory as 'C-states.' It is important to note that C-states do not necessarily form a natural kind; they might turn out to be quite a heterogeneous category. Plausibly, they are heterogeneous, as various states, such as sub-personal states in the cortex and internal states of primitive organisms, might both count as C-states.

3.3 The Limits of Tracking

However, it remains too rigid in other respects. Mendelovici (2013) dubs both causal-informational and teleosemantic theories as tracking theories (understood jointly). Tracking theories, according to her, can handle occasional misrepresentation (like mistaking a distant horse for a cow), but they have difficulty accommodating systematic, reliable misrepresentation. 'According to tracking theories, mental representation is a relation of causation or correlation holding between mental representations and things in the world in content-endowing circumstances, e.g., circumstances in which the tokening of a representation is useful, adaptive, or involves a sufficiently strong causal connection' (ibidem, 2).

Teleological theories can only accommodate unreliable misrepresentations that change across generations (because utility would cause tracking relations to stabilize). But this overlooks the possibility of reliable misrepresentations that are too well-behaved. They're typically useful, adaptive, and involve strong causal connections-exactly the markers tracking theories treat as content-determining. She defines them using an analogy to testing: a test can be reliable (yield consistent

results) without being valid (actually measuring what it claims to measure). SATs are famously stable (in the sense that the same person taking it several times across different conditions, i.e., weeks apart) typically does almost the same, provided they did not prepare. However, SATs fail to predict what they are purportedly measuring: future academic prowess. Clearly these tests measure something reliably, but not veridically. Similarly, a mental representation can be reliable (responds consistently in similar circumstances) without being veridical (accurately representing reality).

This suggests strongly that there has to be a more refined understanding of various types of representational error. This typology should not be limited to incidental/one-off/aberrant vs. systematic/reliable errors, however. Similarly, we should be able to differentiate errors in ascribing properties to judgments about presence or absence. It is one thing to intentionally dwarf the disliked cousin in a family photo (using clever perspective tricks or software) and another to remove him from the picture altogether (again through manipulating the shot or just post-production erasure). In more serious terms, a scientific model that does not represent what it purports to represent (e.g., a model of the atom where the electrons are nowhere to be found) presents a different kind of failure (misrepresenting) than a scientific model that is not isomorphic to its target system (e.g., relative sizes of electrons are distorted).

This is not a banal point, either. This points to the necessity of a pluralistic account of representational error, in which I subscribe to Bielecka's (2025) conclusions in her most comprehensive account of representational error in the literature currently. However, all cases of detecting, reacting, and acting upon representational error share a common kernel. To quote Bielecka again, 'detecting representational errors amounts to detecting discrepancies between batches of information' (Bielecka, 2019:209). We can see here that the hallmark of a

representational system able to misrepresent is a system that has access to at least two (at least partially) independent sources of information on the same target. Also, it is important to notice here, that this does not logically entail the commitment to metarepresentations within the system that are used to judge whether the error occurred in the first place. That would naturally lead to a vicious regress, rendering the theory moot.

The additional condition imposed on that mechanism guarantees that at least one of the compared batches has to be semantic (defined as possession of the satisfaction conditions, i.e., being representational, which licenses normative judgments about the accuracy of received information). For this reason, Bielecka calls her account a 'coherentist theory of representational error.' Organisms are in the business of checking independent sources of information against each other and looking for discrepancies. Whenever these occur, it is a signal for the cognizer to try to assess the conflict and decide which source of information is more reliable (accurate).

4. Back to Front: Forward-Looking Theory of Content

Cameron Buckner's "forward-looking theory of content" is a naturalized philosophical theory of mental content designed to resolve persistent issues in "backward-looking" accounts, like traditional teleosemantics, and to accommodate modern predictive processing approaches to the mind. Buckner introduces his theory by contrasting it with backward-looking teleosemantics (e.g., Fred Dretske's causal informational theory), which fix the meaning or content of a mental representation based on its evolutionary or learning history (its past causal or informational relations).

As we have seen, backward-looking theories are in trouble when faced with explaining how to ascribe determinate content and allowing for misrepresentation (error). To remind, backward-looking views face a dilemma: if the connection is strict (strong correlation), the content becomes so broad that it includes common errors. For instance, if a frog's neural state (N) is historically selected to indicate "fly" (F), but it also gets triggered by a lead pellet, the content of N is forced to be "fly or lead pellet" (or just "small, dark, moving speck"). In this case, the frog can never make a mistake, thus failing to allow for plausible misrepresentation. Simultaneously, when the connection is weak (allowing for error), the content becomes indeterminate or too vague, failing the goal of assigning clear, precise meaning.

Difficulties in clearly and precisely defining what a representation is about, while also failing to explain how a representation can be false or inaccurate (tokened in the absence of the thing it is supposed to represent), are precisely what Buckner's account attempts to correct. To ameliorate that, Buckner suggests looking forward, not backward, in order to fix contents and account for representational error. Thus, he argues that the solution lies in grounding content in the agent's present and future epistemic capacities, rather than its history. The central innovation is that content attributions must be grounded in the agent's own ability to detect when it has represented the world incorrectly through the assessment of prediction errors. After all, representational agents are not passive victims of error; they react to errors by revising their representational schemes to make those errors less likely in the future. And part of having a representation of X is the organism's inherent tendency to treat evidence of tokening that representation in the absence of X as a mistake.

Here his account is similar to Bielecka's, but there are differences, too. For Buckner's core proposal for fixing content is treating content as the 'likeliest

revision trajectory.' But what does it mean for representation's content to indicate at the end of its likeliest revision trajectory? First, the organism's current representational scheme is taken as a starting point. The content is determined not by what caused the representation in the past, but by what the system would converge on if it were to continue learning and revising its representations in response to its own prediction errors. This forward-directed epistemic capacity constrains the space of possible misrepresentation, allowing for errors while maintaining determinate content.

Looking at this proposal, it is not difficult to guess that the forward-looking theory is explicitly designed to integrate with predictive processing with their insistence on cognitive agents being in the business of constant and dynamical monitoring of their own errors and controlling the trajectory of epistemic accuracy towards a more optimized world model. In PP, the mind is seen as a hierarchical system that constantly generates predictions about sensory input, and cognitive work is done by processing the resulting prediction errors. Buckner's theory finds a natural fit here: the prediction signal is the very mechanism by which the agent "detects" a representational error, and the subsequent model updates (revisions) driven by the PE signal determine the final, forward-looking content. The theory thus provides a framework to make content attributions to the representations (models/priors) posited by predictive processing approaches, addressing the philosophical challenges that older naturalized theories of content faced.

What is particularly interesting for our purposes is Buckner's suggestion that within this account of representation, compression (I am expanding on this point significantly in chapter 8; in fact, it is the topic of this entire chapter) plays a crucial role. According to him, compression is central to explaining how mental representations achieve determinate content while remaining useful and adaptive. By defining ‘content’ as what the system converges on through error reduction over

time, the theory implicitly relies on this drive for compression. The representation's final, determinate content is the most compressed, high-level summary of the world's causal structure that still maintains a minimal prediction error. To go back to the famous frog, if a creature's representation is compressed to "small, dark, moving thing," it will generate prediction errors when it tries to act on a lead pellet. The likeliest revision trajectory (the forward-looking content) is the more finely specified, slightly less compressed representation, such as "insects," that successfully minimizes future prediction error and maximizes its utility.

As we know, within the Predictive Processing (PP) framework that Buckner utilizes, the brain's "representations" are essentially generative models that predict incoming sensory data. For these models to be effective, they must be highly efficient; they must represent the world using the least amount of complexity necessary to minimize prediction error. This pursuit of efficiency is a drive toward compressed representations. Successful compression involves ignoring insignificant noise and capturing only the statistically significant, recurring, and relevant causal structure of the environment.

However, one might object that the drive towards increased predictive success and compressed representations needn't necessarily align. In fact, these directives might pull the system toward differing (or even conflicting) directions. The dynamical, forward-looking view of content as gravitating towards the more optimized representation without discarding compressive properties means striking a balance between redundancy compression and predictive differentiation. That is why the system is, in a sense, in constant adaptive flux and the contents are never fully fixed, as the pressure along any of these axes might increase, forcing the renegotiating of this balance. This is precisely the postulated mechanism for error correction, understood both as accommodating one-shot errors and also adjusting to the long-term makeup of the informational environment.

Buckner addresses these worries head-on:

"On the one hand, the drive to correctly predict stimulus patterns at the output layer creates a pressure to differentiate stimulus patterns which predict different outcomes, by altering link weights to render their hidden layer activation patterns more distinct. On the other hand, there are fewer hidden layer nodes than input or output nodes, so the system cannot just memorize every statistical regularity in its environment; it must economize on representational resources by treating input stimulus patterns which redundantly predict the same outcomes as being the same, by rendering their hidden-layer activation patterns more similar. The balance between these two pressures [...] enables a powerful form of all-purpose inductive learning, which over time will tend to discover an efficient set of stimulus configurations that predict the widest range of stimulus variance in the organism's environment" (Buckner, 2021: 382).

Two natural objections come to mind when faced with such a proposal that locates the 'actual' content in the future, intuitively outside the grasp and disposal of the agent purportedly using it. How on earth could the cognizer utilize something inaccessible to it to guide its behavior? How could it cause it? Let's call this 'backwards causation.' But additionally, wouldn't forward-looking content ascription render any non-(yet)-optimal (if one can ever be within a reportedly always-shifting outlook) agent in a sense not good enough, as in 'unable to utilize the actual content that we try to ascribe to it as causing its behavior? Let's call this 'inaccessible content.' Buckner suspects these objections stem from his theory's reliance on future-oriented content. He provides specific rebuttals to each.

First, as mentioned, the objection claims that if a representation's content is fixed by its future optimal state, then the future is causally determining the present state of the representation, which implies backward causation (let’s call it just like that).

To counter ‘backward causation’ criticism, Buckner argues that there is no need for oracular knowledge, which this criticism implies. The point of the criticism is that the organism would need access to a perfect outside vantage point in order to check his predictions against and finally arrive at the definite content that his learning curve would eventually converge on. But, he clarifies that the content is fixed not by the actual future state, but by the forward-directed epistemic capacities (mechanisms for learning and error correction) that the agent currently possesses. That is because the forward-looking claim is about the normative trajectory of the system, not an actual causal influence from a later point in time. Thus, the organism doesn't need "oracular knowledge" of its final, error-free state. Its current state (a representation R has content C because R is currently deployed as if C were its content, and the system is committed to the revision process that, if completed, would result in C. The content is determined by the dispositional properties of the system now-its inherent drive and capacity to detect and correct errors.

Now let's turn to the second charge and his rebuttal to ‘Inaccessible Content’ worries. Again, the objection states that if the content of a representation is the state it will have at the end of its likeliest revision trajectory (i.e., in the future), that content is inaccessible to the current agent, meaning the agent cannot possibly act upon it. Forward-looking theorists' answer here would be that content is currently effective. Buckner agrees that the perfectly revised content is a theoretical endpoint but argues that the current, imperfect content is already effective in guiding behavior. The organism's behavior is explained by the current state of its internal model (R), which is already functionally committed to a particular revision strategy. For example, an agent's current model might predict that an object will behave like a cup. When the object shatters (generating a prediction error), the behavior is not random. The prediction error immediately triggers a revision process that is only

comprehensible by reference to its target: the new, more accurate concept of a cup (or maybe, "fragile cup"). Therefore, the forward-looking content (C) provides the normative framework that makes the agent's current, fallible behavior (including error detection and correction) explanatorily coherent. The agent acts on its current model, and its future content is what makes its current corrections intelligible.

5. The Mechanism of Content: Algorithmic Homuncularism

5.1 Algorithmic Homuncularism as the Normative Frame

We have established that representations must be structurally rich and forward-looking. However, identifying the theoretical profile of content is insufficient; we must identify the causal machinery that fixes it. How does the brain implement the norms required to distinguish a correct representation from an error? To answer this, we must look for an architectural feature that acts as a 'normative anchor'—a mechanism capable of validating information flow. This brings us to the concept of algorithmic homuncularism.

Algorithmic homuncularism (AH), endorsed most notably by Shea (2018), is in its essence a commitment to the project of explanation via decomposition into smaller parts (the ‘stage model’ we encountered in section 1 of chapter 1). In terms of neuroscience, or rather philosophy in/of neuroscience, it means stalwartly subscribing to the idea of compositional, explanatory decomposability (Bechtel & Richardson, 2010).

What is critical is that the postulate of decomposability contains both spatial and temporal divisibility. Parts of the brain should be localizable in virtue of the content they trade in. But also, the distinct steps of the algorithm should be clearly

separable in explanations. And what's more, and more fundamental, the transitions from one step of the algorithm to the other have to be caused by the content encoded by the parts. As we see, the bar is raised high for AH. However, the classical informational project relies heavily on the alethic status of AH. Hence, I strongly believe we should fight for it.

AH proposes that explanations of cognitive function require identifying and localizing computational steps within the physical structure of the brain. The view is characterized by three essential commitments:

1. [Injective Mapping]: Requires that distinct contents be mapped onto spatially and functionally distinct vehicles. What's important, though, is for this mapping to be unequivocal, e.g., targeting not more than one content (hence injection). For every content postulated at a stage in an algorithm, there must be a particular physical part of the system (a vehicle) that realizes that content, and this part must be distinct from the vehicles for other contents or stages. AH requires that vehicles be arranged in a manner that supports this strict localization.

2. [Vehicle Realism]: For reasons above, AH endorses strong vehicle realism, the commitment that there must be real, physical vehicles that realize contents in order for representational explanation to be feasible (Burnston, 2020, 2021). AH posits that representation arises where a system implements an algorithm for performing task functions, meaning that internal vehicles bear exploitable relations to features of the environment and are processed internally in a way appropriate to those relational properties.

3. [Causal Isomorphism]: The connection between spatial and temporal localization necessitates that the causal interactions of these distinct

> physical entities be clearly defined and discrete, reflecting the exact steps and content transformations proposed within the algorithm. The causal steps must process the contents in accordance with the transformation described by the algorithm.

With the architectural necessity of the consumer established, we have solved the structural problem: there is a mechanism capable of validation. However, this only tells us *that* the system selects content, not *which* content it selects. We still face the problem of indeterminacy: among the infinite correlations a neural signal might carry (e.g., 'fly,' 'black dot,' 'nutritious object'), which one constitutes its genuine semantic content? To resolve this, we must equip our Algorithmic Homunculus with a specific selection criterion—a rule for identifying the signal that matters. This brings us to the concept of unmediated explanatory information."

5.2 The Normative Anchor: Why AH Matters

But why is this architectural commitment crucial for a theory of content? The answer lies in the normative role the algorithm plays. We cannot determine what a representation refers to until we identify the algorithm the system is executing. The algorithm defines the consumer's need. The "algorithmic homunculus" serves as the normative anchor—it sets the standard for what counts as a "correct" input, thereby fixing the semantic domain.

Central to this normative role is the concept of the exploitable relation. The content is not just any correlation; it is the specific relation in the world that the algorithm is built to exploit to achieve success. For instance, a navigation system might exploit the relation between "polar angle" and "movement vector." The algorithm's structure—its specific computational steps—reveals that it is "looking for" angular information, not Cartesian coordinates. Thus, the algorithm itself helps determine

the content by defining which environmental features are computationally relevant for the system's success.

5.3 Unmediated Explanatory Information as Content Filter

But even with a localized vehicle and a defined algorithm, we still face the problem of content indeterminacy—identifying exactly what that vehicle represents among the many things it correlates with. To solve this, we borrow the idea of unmediated explanatory information from Shea (2018). This concept helps us make progress on the problem of content indeterminacy by specifying which correlations between a neural vehicle and the world are genuinely content-determining. Its goal is to fuse elements of teleosemantics (function) and informational semantics (correlation) while constraining the latter with an explanatory requirement.

In his book, Shea offers the following formal definition for unmediated explanatory information:

'The UE information carried by a set of components R_i in a system S with task functions F_j is the exploitable correlational information carried by the R_i which plays an unmediated role in explaining, through the R_i implementing an algorithm, S's performance of task functions F_j' (Shea, 2018:84).

This definition is used to ground his general condition for content based on correlational information: "If component R of a system S with a task function or functions F carries UE information about condition C, then R represents C." (Shea, ibidem).

The central concept to his account is the distinction between a mediated and an unmediated explanatory role. This distinction is what allows Shea's theory, which he calls ‘varitel semantics’ (a wordplay on 'variety of telos'), to pick out the correct

distal content and solve the indeterminacy problem. Shea clarifies this distinction using the famous example of the frog's fly-catching mechanism:

'The idea that some correlations play an unmediated role in an explanation calls for clarification. In the classic example of the frog's fly-catching mechanism, the correlation of retinal ganglion cell firing (R) with little black things (C) figures in an explanation of how the system was stabilized by evolution, but that explanation also mentions the fact that being a little black thing (condition C) correlates with being a yummy flying thing (condition C'). Without that background correlation, it would be unclear how the correlation between R and C enabled frogs to achieve an evolutionarily beneficial outcome. So the role of the $R-C$ correlation in that explanation is mediated. There is another explanation of stabilization that adverts directly to the correlation between R and nutritious flying objects (C'). The role of the $R-C'$ correlation in that explanation is unmediated' (Shea, ibidem).

He then gives the general principle for the distinction: 'A correlation between an item R and condition C plays a mediated role in an explanation if its role depends on the explanation adverting to a further correlation between C and some further condition C'; otherwise it plays an unmediated role.' (Shea; ibidem).

Crucially, this addresses the robustness or correlational stability required for content. The system doesn't just track any correlation; it tracks the one that is robustly linked to survival or task success across the relevant range of environments. The "little black thing" correlation is fragile; it fails in environments with lead pellets. The "nutritious object" correlation is what explains the persistence of the mechanism. Thus, UE information acts as a semantic filter of sorts: it screens out proximal cues (black dots) and selects the distal condition (food) that makes sense of the behavior in the light of evolution. This provides a tentative but convincing (I think) answer to the indeterminacy problem raised by Quine and Cao:

the content is fixed not by the raw signal, but by the explanatory logic of the system's survival.

5.3.1 Format and Immediacy: Deepening the UE Picture

This picture of UE information can be further clarified by connecting it to Kulvicki's analysis of representational format. Shea fuses teleosemantics with information theory to solve the problem of content indeterminacy. The crux is that we must identify which of the many correlations carried by a vehicle plays an unmediated role in explaining the system's success. Recall the frog. Its retinal cells correlate with "little black things" (proximal) and "nutritious flying objects" (distal). Both correlations exist. But the functional explanation of why frogs have this mechanism must appeal to nutrition, not blackness. The blackness is merely a cue. Thus, the correlation with "nutritious object" plays the unmediated explanatory role. In formal terms, the information about nutritious objects is what is explicit or immediate for the frog's feeding behavior system, even though information about blackness is also present (and indeed is what the mechanism actually detects in the sensory periphery).

Kulvicki's three components of immediacy—extractability, syntactic salience, and semantic salience—can be understood as specifying what makes information unmediated:

1. The R−C′R-C' R−C′ correlation must be extractable—there must be nonsemantic features of RR R that carry information specifically about nutritious flying objects (via the reliable environmental correlation with little black things).
2. The features of RR R that carry information about C′C' C′ must be perceptually/computationally salient to the frog's downstream systems (syntactic salience).

3. The plan of correlation between RR R and C′C' C′ must be one that the frog's feeding system can exploit—the frog doesn't need to "know" that black things indicate nutrition; its circuitry simply needs to implement an algorithm where RR R-states trigger feeding behavior (semantic salience).

The crucial point is that Shea's account explains why these features matter: they matter because evolution has shaped the system so that information at this particular level of abstraction (nutrition, not blackness) plays an unmediated role in explaining task performance. This deepens our understanding of why format is crucial for the entire project of informational explanations of cognitive systems. The format determines what information is immediately available to downstream processes—and this immediacy is precisely what UE information tracks.

5.4 The Empirical Challenge: Algorithmic Coherence

However, AH faces challenges, particularly from those who argue that the complexity observed in neural population responses prevents the localization of distinct contents to spatially and temporally distinct parts of the system. This critique often motivates an alternative view, Algorithmic Coherence (AC), which suggests that brain processes may correlate with algorithmic transformations even if discrete contents cannot be assigned to distinct brain areas. AC claims, conversely, that 'representations are realized as trajectories in the state space of a whole population, rather than as isolated to distinct parts of that population. "Hence, injective mapping is false. [...] What is important about this, for current purposes, is that any measure of representation of a particular variable is taken to be constituted by the whole population, rather than by distinct parts within that

population. If so, then the system cannot be divided in the way AH recommends"[53]' (Burnston, 2021: 1620, 1624-1625).

These types of views were emboldened by the groundbreaking study by Mante and colleagues (2013). The study seemed to demonstrate that in the face of modern, more thorough experimental protocols, the received, venerable picture of stable selective receptibility to a particular type of stimulus does not hold water anymore, even for single neurons. The central purpose of the Mante study was to examine PFC activity in monkeys trained to flexibly select and integrate noisy sensory inputs (motion and color) based on a specific behavioral context, ultimately leading to a choice. Their results pointed not to stable classical encoding properties but towards distributed and complex encodings. The activity of individual prefrontal neurons was remarkably complex, generating responses that defied deep understanding if analyzed in isolation. In turn, the researchers found that information pertaining to motion, color, context, and choice were all distinguishable at the population level of activation.

Also, the neurons in the population exhibited widespread 'mixed selectivity' (Rigotti et al.,2013). This means that rather than being specialized for a single experimental variable, individual cells were selective for multiple task parameters simultaneously. They interpreted their result as that the observed complexity and the functional roles of single neurons were best understood only when viewed through the framework of a dynamical process unfolding at the level of the

---

[53] It is important to note that the debate between algorithmic homuncularism and algorithmic coherence regarding population trajectories (Burnston vs. Kinsey) is explicitly a debate about the *format* of the neural vehicle (specifically, its spatial and temporal decomposability). However, we treat it here in the context of *content* because this physical separability is the necessary condition for the emergence of the 'consumer'—the normative anchor required to fix semantic content. Without this structural decomposability, the distinct processing stages required to validate content cannot be distinguished from the continuous flow of dynamical activity, and the 'view from nowhere' required for determinate content collapses. Thus, population coding is analyzed here only insofar as it grants us the license to ascribe content.

population. The mechanism responsible for both the selection of relevant input and the integration of that input were determined to be two aspects of a single dynamical process unfolding within the same prefrontal circuits. This suggested a previously unknown, holistic mechanism for context-dependent computations.

It is not difficult to see whether it seems to spell trouble for classical view and algorithmic homuncularism and provide critical evidential support for the population doctrine. Additionally, it seems to also undercut injective mapping and causal isomorphism and vehicle realism. In the population picture, in place of stable mules of meaning, we encounter chameleons—non-distinct vehicles. The analysis indicated that the principal components (or 'task axes') corresponding to each task parameter—motion, color, context, and choice—were weighted positively for the entire population over the duration of the task. And here comes the crucial catch - this means that the representation of the action choice was superimposed upon representations of motion and color across all neurons in the population at each stage of activity.

Danger to vehicles is also detrimental to decomposition. Since any measure of representation for a particular variable was constituted by the whole population, rather than isolated to distinct parts of it, the system cannot be divided in the strict way AH recommends. The system targeted by Mante et al. was characterized as a non-decomposable system where there is "no way to divide the system into spatially and functionally distinct components" (Kinsey, 2023). But to make matters purportedly worse, AH requires distinct contents to be tokened at temporally distinct processing stages, connected by discrete causal transitions. The dynamical nature of the PFC response challenged this premise, too. Because information was superimposed and present throughout, there was no clear temporal stage at which one content was tokened and the others weren't tokened yet. Nor were there "clear phases of transition between some represented contents and others" (Burnston,

2020). This opaqueness of the process makes it difficult to describe it in terms of the clear, sequenced stages required by algorithmic analysis.

5.5 Defense: Kinsey's Temporal Decomposition

However, Kinsey (2023) demonstrates a convincing way forward, impressively digging even deeper in papers building upon Mante's original study. The battlefield is the temporal dynamics and its purported decomposition into semantically separate tokens. As we remember, Burnston claimed that systems like the prefrontal cortex (PFC) epitomize non-decomposable systems. Kinsey argues that a follow-up analysis of key empirical data demonstrates that the representational properties of the system do not remain static over the task duration. He cites Aoi et al.'s (2020) study, where researchers performed dimensionality, allowing them to analyze fluctuations in how components captured variance over time. Kinsey interprets their results as showing clear temporal sub-regimes or epochs, providing strong evidence that while the system's representational properties are not static, they can still support the temporal division necessary for algorithmic sequencing.

Additionally, while Burnston cited Mante et al. to support the critique of AH due to mixed selectivity, Kinsey engages with this specific study to argue that it can be successfully modeled within Shea's framework. Kinsey uses the findings that motion, color, context, and choice information were distinguishable in the population to illustrate how the application of UE information resolves the ambiguity of mixed selectivity, thereby defending the interpretation of the system's sequenced computation. Finally, he demonstrates that recent developments in theoretical neuroscience concerning the dimensionality and decodability of neural representations in fact do in practice constrain researchers's hypotheses about representation. He argues that decodability constrains whether UE information can

be ascribed to various stages of activity. This literature suggests that information may be present (decodable) but in a "badly formatted" way for downstream use during early stages. This insight supports the idea that even if representations overlap in the state space, the information may not be functionally available or exploited until later, distinct stages of processing, thereby constraining the content assignment and undermining Burnston's claim. An advocate of AC hence finds herself caught on the horns of the dilemma: if information is not available during early stages and yet is exploited, she must either concede that it is exploitable during later stages (thus conceding the sequential stage-like process) or deny that information is available at all.

Establishing the existence of a consumer mechanism solves the internal problem of validation, but it leaves the external problem of *grounding* untouched. How does this internal coherence relate to the external world? The answer is that the algorithmic homunculus does not operate in a vacuum; it operates within a specific environment governed by stable causal laws. The 'burst' of validation is not merely a sign of neural synchrony; it is the biological registration that a specific environmental constraint has been satisfied. To understand how these physical checks map onto semantic content, we must turn to the logic of information flow itself—specifically, the framework of situation semantics.

5.6 Integrating Situation Semantics: Constraint-Based Content

Thankfully, as a result, we can deepen this account by integrating it with Israel and Perry's (1990) situation-theoretic account of information, which we luckily introduced in Chapter 1. The integration of Israel and Perry's constraint-based information theory with Shea's teleosemantic account of content determination reveals deep structural compatibilities that illuminate longstanding puzzles about representational format and mental content.

Shea's concept of unmediated explanatory information provides exactly the selection mechanism needed to determine which of the many correlations carried by a vehicle constitute its genuine representational content. Where Israel and Perry's framework recognizes that a single fact can carry information about multiple conditions relative to different constraints, Shea's criterion specifies which correlation is content-determining: the correlation that plays an unmediated role in explaining the system's performance of its evolutionarily or developmentally selected task functions. Shea's teleosemantics thus provides the naturalistic grounding for constraint selection that situation semantics requires: evolution and learning determine which constraints systems become attuned to by selecting for systems that exploit fitness-relevant correlations.

Crucially, this integration resolves the apparent tension between many-to-one mappings in information theory and the requirement for determinate content in teleosemantics. When Israel and Perry claim that multiple vehicles can carry the same informational content, they are addressing the content-to-vehicles direction. Shea's project, by contrast, addresses the vehicle-to-content direction: given a particular vehicle V (such as retinal ganglion firing), which of its many correlations constitutes its representational content? His answer—the correlation playing an unmediated explanatory role for task functions—makes this mapping determinate or at least strongly constrained for each system-task pair. These are not contradictory requirements but complementary aspects of a complete theory.

The notion of receiver relativity or consumer specificity emerges naturally from this synthesis. A representational vehicle carries information in a particular format "for" a downstream consumer system with specific task functions and constraint attunements. The same neural population activity might carry spatial information in polar format for motor control systems attuned to constraints linking angle-distance parameters to action affordances, while simultaneously carrying

information in Cartesian format for spatial memory systems attuned to constraints linking orthogonal coordinate structures to allocentric representations.

Situation semantics provides the abstract logical structure for constraint-based content, explaining *that* information depends on regularities. However, it does not explain *how* a biological system, constrained by metabolic limits and lacking a 'view from nowhere' of these regularities, can physically implement these checks. We need a mechanism that converts abstract constraints into concrete causal forces. This requires us to return to the biophysical architecture of the two-point processor (introduced at the end of Chapter 3) to see how it functions not just as a format, but as a semantic engine.

6. The Biophysical Implementation: The Two-Point Processor

6.1 The Mechanical Judge: Solving Semantic Smuggling

Recent attempts to mechanize the interactivist framework, such as those by Miłkowski (2025), posit that representational error can be naturalized through a process of inconsistency detection (inspired by Bielecka's 2019 coherence-based account of error). In this view, the cognitive system validates its representations by comparing multiple independent information streams - for example, comparing sensory evidence against internal predictions or cross-referencing different sensory modalities. If the streams diverge, an error signal is generated. While this approach ostensibly avoids the need for an external interpreter to verify truth, it risks reintroducing a subtle form of homuncularism at the architectural level.

The difficulty lies in the preconditions of the comparison itself. For a system to detect an inconsistency between Signal A and Signal B, it must arguably possess a

priori knowledge that Signal A and Signal B share a referent. If the system compares a visual signal of a vertical edge with an auditory signal of a loud pitch, their divergence is meaningless unless the system already knows these signals should be correlated. If this semantic equivalence is hardwired or assumed as a premise of the architecture, the theory smuggles in the very intentionality it aims to explain. This constitutes a petitio principii: the semantic content (the knowledge of reference) is treated as an input to the error-detection mechanism rather than its output.

To resolve this circularity, we must identify a mechanism that can adjudicate between information sources based solely on their structural and causal properties, without requiring semantic insight. The comparison must be physical, not logical.

6.1.1 The Two-Point Processor Architecture

The two-point processor model provides the necessary biophysical architecture to solve this problem. Unlike the standard point neuron, which sums all inputs into a single variable, the pyramidal neuron maintains a strict structural asymmetry between two distinct integration zones: the basal dendrites and the apical dendrites. This separation allows the neuron to distinguish mechanically between driving inputs (content) and modulating inputs (context).

Importantly for error detection, this architecture replaces the abstract comparator with a concrete biophysical operation: coincidence detection. The neuron does not compare the semantic content of its inputs; it detects the temporal coincidence of their arrival. The basal tree receives feedforward sensory data, providing the specific receptive field information (e.g., vertical edge). This anchors the representation to the external world. The apical tree receives feedback and lateral input, providing the contextual field (e.g., expectation or cross-modal confirmation).

Crucially, these inputs interact non-linearly. The apical input is asymmetric and modulatory; it cannot drive the cell to fire on its own. Instead, it alters the gain of the cell's response to the driving input. This structural constraint ensures that the system does not hallucinate context as content[54]. The knowledge that Source A and Source B are relevant to each other is not represented as a semantic rule but is instantiated by the physical connectivity and the requirement for temporal synchrony (~15 ms window).

Bursting as Validation. The output of this coincidence detection is a phase transition in the neuron's firing mode, mediated by back-propagation activated calcium (BAC) firing. This mechanism serves as the mechanical judge of coherence. When the driving input (content) matches the modulating input (context), the coincidence triggers a calcium spike in the apical trunk. This calcium event converts the neuron's output from a single action potential into a high-frequency burst (typically 2–4 spikes). This burst is the physical token of validation. It signals to downstream consumers that the feedforward content has been verified by the internal context. Conversely, if the inputs are inconsistent—or if the context is absent—the calcium spike is not triggered, and the neuron fires only a single spike (or is silenced by inhibitory veto mechanisms).

This mechanism effectively gates information based on its coherence. The system does not need to know that the inputs refer to the same object; it only needs to be tuned to propagate bursts while filtering out single spikes. Thus, the consumer of the representation is not a homunculus reading a code but a downstream synaptic chain that responds selectively to the high-intensity burst signal. This converts the abstract problem of error detection into a concrete problem of signal amplification: coherent (valid) signals are amplified into bursts, while incoherent (invalid) signals

[54] Marvan et al. (2021) represent the philosophical extension of this biological program. They argue that Apical Amplification is not merely a computational mechanism for context-sensitivity, but the specific "conscious-status awarding mechanism" (CSAM) required for subjective experience.

are attenuated, solving the problem of semantic smuggling through purely causal mechanisms.

Content is the Base, Context is the Apex. To fully appreciate why this architectural asymmetry resolves the homuncular regress, consider the difference between a driver stopping at a stop sign and a driver stopping because their vehicle has impacted a concrete barrier. In the first instance, the cessation of movement is an act based on semantic interpretation; the driver must perceive the sign, understand its symbolic meaning, and voluntarily apply the brakes. In the second instance, the cessation of movement is an act based on physical force; the barrier mechanically prevents forward motion regardless of the driver's understanding or intent. The fundamental challenge for a naturalized theory of mental content is to explain how a semantic property—such as the coherence between a sensory input and an internal expectation—can acquire the causal efficacy of a physical barrier without requiring an intermediate interpreter to read the sign and decide to act.

Phillips' framework achieves this naturalization by converting semantic coherence into biophysical force. In this model, the contextual input arriving at the apical dendrites functions as the sign. It carries information about the broader structural regularities of the environment, such as the prediction that a vertical edge should be present given the surrounding visual features. However, unlike a sign that must be interpreted, this contextual information is coupled to a biophysical mechanism—apical amplification—that acts with the efficacy of a barrier or a gate. When the context matches the content carried by the basal dendrite, the neuron undergoes a phase transition from a weak spiking mode to a high-intensity burst mode. This burst is not a symbol that requests attention from downstream systems; it is a biophysical force that physically compels downstream synapses to fire via the mechanism of short-term synaptic facilitation. Thus, the system does not need to

know that the inputs are relevant; the relevance is instantiated by the physical synchronization of the inputs, which unlocks the gate to downstream propagation.

This distinction clarifies why the emergence of the two-point processor represents a qualitative shift in neural computation rather than a mere quantitative increase in complexity. The interaction is fundamentally asymmetric. The basal dendrites provide the driving input that defines the content—the what of the representation. The apical dendrites provide the modulating input that defines the gain—the how much of the transmission. This asymmetry ensures that the context does not contribute new features to the content but rather acts as a meta-operator. It determines the salience of the signal without altering its semantic payload. This vindicates the intuition that modulation is distinct from contribution; the context facilitates the transmission of the content but remains distinct from the content itself, preserving the objective reference of the signal while enabling its subjective prioritization.

Furthermore, this framework forces a reconsideration of the role of suppression in neural processing. In the context of Phillips' 'cooperative neuron,' suppression is an active implementation of content specification. By suppressing incoherent signals—those that do not match the prevailing context—the system effectively purifies the content stream. This operation can be understood through the distinction between filtering and selection. Selection requires the active amplification of a target, a process energetically expensive and implemented via the burst code. Filtering, conversely, is a passive but critical operation implemented by the inhibitory veto of somatostatin-positive interneurons. By systematically removing noise and irrelevant data, the system ensures that what remains is not just a raw sensory feed but a curated representation of reality. Thus, suppression is not a failure to represent; it is a necessary operation that defines the boundaries of the content that is eventually consumed.

The verdict of this analysis is that the two-point processor provides the essential mechanism for a naturalized account of semantic content. A representational format, such as a vector code, is insufficient on its own to explain why certain neural states drive behavior while others do not. For content to be causally efficacious, there must be a mechanism that distinguishes between a latent signal and a verified command. The two-point architecture serves as this causal switch. It allows the brain to mechanically distinguish between signal (validated content) and noise (unvalidated content) without appealing to a homunculus. The result is a system that behaves for reasons—specifically, the coherence between internal models and external data—but operates entirely by physical causes. Representation, in this view, is the mechanism of amplifying shared structure while suppressing unique noise, effectively operationalizing content as that which survives the coherence check.

6.1.2 The Biophysical Basis of Error Detection

A potential objection to this proposal is that replacing a point neuron with a two-point processor merely shifts the problem of the consumer down a level. If the neuron acts as a comparator that outputs a burst code, does this not simply require another, more complex consumer to interpret the burst? If the downstream system must know that a burst means validated content, have we not just reintroduced the homunculus at the synapse?

This objection fails because the reading of the burst code is a constrained biophysical reaction, not a semantic interpretation. Phillips identifies short-term synaptic facilitation as the mechanism that tentatively terminates the regress (2023: section 5.5.3). Downstream synapses are physically hardwired to respond reliably only to high-frequency trains of action potentials (bursts) while remaining unresponsive or depressed in response to low-frequency single spikes. This tuning is a property of calcium accumulation in the presynaptic terminal, a blind physical

process that requires no semantic insight. The downstream neuron does not decide to accept the burst; it is physically compelled to fire by the non-linear increase in neurotransmitter release. As Phillips notes, the output bursts of two to four action potentials produce short-term synaptic facilitation, which greatly strengthens the cell's effect on the sites to which it projects. Thus, the infinite regress is halted by the conversion of an informational state (coherence) into a raw causal force (facilitation).

Identifying the mechanical judge explains how the system validates a signal, but it does not yet explain what objective function this validation serves. If the burst is the token of success, what computational goal is the system optimizing by producing it? The answer cannot be simple information maximization, which would leave the system vulnerable to high-entropy noise. Instead, the two-point processor implements a more sophisticated algorithm designed to maximize coherence.

6.2 Validation via Coherent Infomax & Precision

6.2.1 The Computational Objective: Coherent Infomax

Having established the biophysical hardware of the two-point processor, Phillips defines the computational algorithm that this architecture implements: coherent infomax. He contrasts this with standard infomax principles, such as those proposed by Linsker, which operate on the assumption that the goal of sensory processing is to maximize the total mutual information between input and output. Phillips argues that standard infomax is insufficient for biological survival because it is blind to the quality of the correlation. A system optimizing standard infomax would treat the faithful transmission of high-entropy noise—such as random static—as high information transmission, potentially overwhelming downstream processing with irrelevant data.

Instead, the neocortex optimizes coherent infomax, which seeks to maximize the information transmitted about the receptive field that is coherently related to the contextual field. Mathematically, this is expressed as maximizing the three-way mutual information between output, context, and receptive field input, under the constraint that the contextual field modulates the output without corrupting the semantic meaning derived from the receptive field.

To rigorously quantify this, Phillips employs partial information decomposition (PID). PID decomposes the total mutual information into three distinct components: unique information (information present only in the driving input, often corresponding to sensor-specific noise), shared information (information present in both the driving and contextual inputs), and synergistic information (information available only by knowing both inputs together).

The biological operation of apical amplification acts as a filter designed to maximize the transmission of shared source information. By amplifying the signal only when the basal input (content) matches the apical input (context), the neuron filters out the unique information (noise) and propagates the shared structure. Crucially, Phillips distinguishes between source synergy (which arises from correlations in the external world) and mechanistic synergy (which arises from the internal operation of the neuron). Coherent infomax utilizes mechanistic synergy—the non-linear interaction of the two input streams—to extract and amplify the source-shared information, ensuring that the system prioritizes purposeful objectivity over raw data throughput.

6.2.2 From Mechanisms to Operations: A Computational Taxonomy

The two-point processor operates through discrete biophysical modes—summarized in Table 3 below. These modes (amplification, attenuation, normalization, isolation, drive) constitute the physical primitives of the system: the

"hardware states" available to the neuron. However, to understand their cognitive utility, we must examine the computational operations these mechanisms implement at the network level. Phillips identifies how these biophysical modes, applied in coordinated patterns, generate specific information-processing functions:

Disambiguation (implemented via selective **amplification**) occurs in noisy or ambiguous sensory environments, where the driving input alone may be insufficient to resolve the identity of a stimulus. The neuron uses contextual input to select one interpretation over another. For example, an ambiguous visual edge may be interpreted as part of a specific contour if lateral apical connections from neighboring neurons signal collinearity.

Grouping (figural synthesis) (implemented via mutual **amplification**) is a constructive outcome of cooperative amplification. When a population of neurons representing aligned features (e.g., collinear line segments) mutually amplify each other via lateral apical connections, they perform grouping. This creates a pop-out effect where the coherent figure is distinguished from the incoherent background noise. This operation effectively binds features into a coherent whole without requiring a central binder.

Divisive normalization (implemented via competitive **attenuation**), mediated by PV interneurons targeting the basal region, scales the output of a neuron relative to the activity of its neighbors. This ensures that the rich get richer—strong or coherent signals dominate the local circuit—and prevents saturation of the network.

Salience control (implemented via asymmetric **amplification**) distinguishes context-sensitive amplification from standard multiplicative gain control. While gain control often implies a symmetric multiplication of inputs, apical amplification allows a weak driving signal to be amplified if the context is strong but prevents the

context alone from driving the cell (in the waking state). This preserves the distal tracking nature of the representation while modulating its impact on downstream processing.

These computational operations emerge from the coordinated deployment of five fundamental biophysical modes, which constitute the complete repertoire of the two-point processor. Table 3 maps the hardware-software relationship explicitly:

Hardware mode	**Biophysical mechanism**	**Software operations**	**Function (information theoretic)**
Amplification	Apical input + basal input ↛ burst (BAC firing)	Disambiguation, grouping, salience control	Increases gain of coherent signal; extracts shared information (synergy)
Attenuation	SST inhibition of apical dendrite	Divisive normalization (with PV), context veto	Suppresses incoherent signals; filters unique noise
Normalization	PV inhibition of soma/basal dendrites	Divisive normalization, competition	Relative scaling; rich-get-richer

			dynamics; prevents saturation
Isolation	Downregulation of coupling (Ih channels open)	Offline processing, raw data mode	Functional decoupling; pure feedforward processing (slow-wave sleep)
Drive	Super-threshold apical input (high ACh, low NA)	Generative simulation, mental imagery	Internal generation; context becomes content (REM/dreaming)

Table 3: Taxonomy of Operations in the Two-Point Processor Operation

The table clarifies the distinction: amplification, attenuation, and normalization are online modes that enable moment-to-moment computation during waking processing, while Isolation and drive represent state-dependent reconfigurations of the processor itself—explored in detail in Section 6.2.3. The software operations (disambiguation, grouping, etc.) are emergent properties of these hardware modes deployed in specific patterns across neural populations.

6.2.3 State-Dependent Dynamics: Online vs. Offline Processing

The efficacy of the mechanical judge is not constant; it is state-dependent. Phillips argues that the brain switches between distinct processing modes based on the level of neuromodulation, which regulates the coupling between the apical and basal compartments. Larkum et al. (2022) identify this active coupling not merely as

a functional feature but as a regulated state that distinguishes conscious processing from the 'functional disconnection' of sleep, which regulates the coupling between the apical and basal compartments. This switching between 'online' (coupled) and 'offline' (isolated/generative) modes implements the 'Cognitive State Matrix' introduced in Chapter 2 (Section 5.3). Specifically, the 'apical isolation' state corresponds to the 'sensory inertia' quadrant, while 'apical drive' provides the mechanism for 'genuine cognition' and simulation.

Online (coupled) processing occurs in the active waking state, where high levels of neuromodulators such as acetylcholine (ACh) and noradrenaline (NA) facilitate the coupling between the apical tuft and the soma. This biochemical state closes leak channels (such as Ih), allowing the back-propagating action potential to reach the apical dendrites and trigger the calcium spike. This is the online state, where the system is engaged with the world and capable of coincidence detection. It is only in this state that the mechanical judge is active, validating content against context.

The specific neurochemical basis for this state has been further elucidated by recent models. The mechanism for this switch is identified in the GANE model (Mather et al., 2016; glutamate amplifies noradrenergic effects), which posits that noradrenaline does not simply arouse the brain globally but amplifies 'local hotspots' of high-priority excitation. Additionally, acetylcholine has been shown to enhance the excitability of somatostatin interneurons, which paradoxically disinhibits the apical dendrite by suppressing parvalbumin-positive cells, thereby permitting compartmentalized dendritic spiking during high arousal.

Offline (isolated) processing occurs during slow wave sleep (SWS) or states of low arousal, when neuromodulatory drive decreases. This leads to the opening of Ih and other leak channels, which effectively short-circuit the dendritic trunk. This results in apical isolation, where the apical dendrite is functionally decoupled from the

soma. In this state, the neuron reverts to a simple point processor driven only by feedforward input. This corresponds to a state of cognitive blindness, where the system processes sensory data (housekeeping or raw input) but lacks the internal model check required for conscious awareness or validation.

This physiological disconnection is corroborated by Gent et al. (2018), who characterize it as a 'functional disconnection' of the cortex. During slow-wave sleep, the shift in thalamic firing modes from burst to tonic, combined with the loss of neuromodulatory drive, effectively amputates the apical context from the basal content, enforcing the isolated state.

Consequently, Phillips's model suggests that the brain's 'offline' state is not merely a passive lack of energy but an active decoupling of apical and basal streams. This decoupling enforces a functional silence—a protective isolation that prevents the system from being overrun by the entropic flood of disorganized inputs. Without this capacity to silence the mismatch, the system would not settle into rest but into delirium: a state of high-firing noise where everything connects to everything else.

Offline (generative) processing occurs in REM sleep (dreaming) or during mental imagery, where a specific mix of neuromodulators (high ACh, low NA) alters the threshold for calcium spikes, allowing the apical input to become driving. This is apical drive. In this state, the context (internal model) generates the content directly, driving the sensory cortex to see things that are not externally present. This provides a biophysical mechanism for generative simulation and recall.

This distinction offers a compelling structural analogy—and potentially a homology of failure modes—to the Bar-Hillel-Carnap paradox. In the semantic theory of information, a contradiction is technically 'maximally informative' (as it excludes all possible states) yet remains semantically void. Similarly, unconstrained neural 'noise' represents a high-entropy state that is informationally dense but

functionally meaningless. Both systems suffer from a specific pathology where unchecked connectivity (or total exclusion) leads to a collapse of meaning. In this light, Gent's 'silence' serves as the necessary constraint—the logical amputation—that prevents this collapse and preserves the system's integrity.

6.2.4 The Predictive Processing Debate: Error vs. Precision

The integration of predictive processing (PP) into cognitive science has sparked a foundational debate regarding the scope and nature of neural representation. The standard formulation of PP posits that the brain minimizes variational free energy by constantly generating top-down predictions to cancel bottom-up sensory inputs. In this view, the primary currency of the neocortex is the prediction error—the residual difference between expectation and reality.

This operational premise has profound metasemantic implications. Mark Sprevak (2019) argues that if the primary vehicles of cortical transmission are error units, this necessitates a fundamental shift in our understanding of representation. In classical cognitivism, a neuron fires to represent a feature (e.g., "there is a vertical edge"). In standard PP, a neuron fires to represent a deviation (e.g., "there is more 'vertical edge' than expected"). Sprevak suggests this introduces a dualism in meaning: the system no longer tracks the world directly but tracks its own "surprisal" or failure to predict. This creates a "two meanings" problem where the semantic content of the brain becomes untethered from the distal environment, trapped in an internal game of error minimization.

Conversely, Rosa Cao (2020) argues that the novelty of PP is largely a "gloss"—a terminological re-description of mechanisms that ultimately serve the same old function: tracking distal targets. She contends that even in PP, the "expectations" are merely auxiliary mechanisms designed to tune the sensitivity of the trackers. The goal of the system is still to represent the world, not just its own errors. For

Cao, the semantic content remains the "old idea" (distal realism), even if the "new labels" (priors/errors) suggest a radical departure. Note that Cao's critical position should not be conflated with Bickhard's or Egan's charges against computational superficiality discussed in Chapters 1 and 2.

Phillips' biophysical framework offers a decisive adjudication of this debate. The "shift" view articulated by Sprevak relies on a specific biological assumption: that the mechanism of prediction is subtractive—that top-down signals cancel out matching bottom-up signals to produce silence (zero error). Phillips demonstrates that for neocortical pyramidal neurons, this assumption is false. Marvan and Phillips (2024) demonstrate that subtractive coding is neurobiologically implausible for neocortical pyramidal cells, which empirically supports the precision-weighting interpretation over error-transmission models.

In the two-point processor model, the apical dendrite (carrying the prediction/context) does not inhibit the basal dendrite (carrying the sensory input). Instead, it targets the calcium integration zone to perform apical amplification. When the internal model (context) successfully predicts the external data (content), the neuron does not go silent; it undergoes a phase transition into a high-frequency burst. Thus, the biological output of a "successful prediction" is not zero; it is a maximal signal. This empirical reality undermines the premise of Sprevak's "shift." The brain does not transmit errors; it transmits validated content. The representation remains anchored to the feature detected by the basal dendrite, supporting Cao's realist "gloss."

6.2.5 Apical Amplification as Precision Weighting

If the brain does not transmit errors, how do we reconcile Phillips' model with the mathematical success of predictive processing? The bridge lies in the concept of precision weighting. In standard PP, "precision" refers to the statistical reliability of

a signal. Attention is often modeled as the optimization of precision—increasing the gain on reliable error units to ensure they update the model.

Phillips fundamentally reinterprets this. He aligns apical amplification directly with the mechanism of precision weighting rather than prediction generation. In this view, the "context" signal arriving at the apical dendrite acts as the precision estimate. High precision (match): if the context confirms the input, the "precision" (apical gain) is high. The cell bursts. Low precision (mismatch): if the context is absent or incoherent, the "precision" is low. The cell fires a single spike or is attenuated.

This interpretation is supported by Auksztulewicz et al. (2017), who demonstrate that "contextual expectancies" in the brain are functionally equivalent to "manipulating the precision" of neural responses rather than just canceling them. The apical dendrite provides the weight; the basal dendrite provides the value. The downstream consumer therefore receives precision-weighted content—a feature representation that has been stamped with a high-confidence marker (the burst).

This "precision-weighted content" supports the move toward action-oriented models described by Tschantz et al. (2020) and philosophically interpreted by Gładziejewski (2015). They argue that biological agents do not aim for a veridical "view from nowhere" but selectively model those aspects of the environment relevant for action. Phillips' mechanism implements this by using precision as a filter for relevance. The "precision" signal (apical input) is not just a statistical variance estimate; it is a relevance vector determined by the agent's goals and history. By amplifying only those signals that match the relevant context (purpose) while preserving their basal coupling to the environment (objectivity), the neuron achieves "purposeful objectivity." It solves the "dark room" problem not by seeking

errors, but by seeking coherence—active opportunities for interaction where the internal model and external world align to trigger a burst of causal efficacy.

The integration of Phillips' biophysics with the metasemantics of Sprevak and Cao leads to a clear verdict. The "shift" to error coding is a mathematical abstraction that does not survive biological scrutiny. The brain operates on precision-weighted content. By reinterpreting "prediction" as "apical priming," we can retain some of the computational benefits of the PP framework (using context to optimize processing) without abandoning the realist commitment that our neurons represent the features of the world, not just the errors of our models—an Achilles heel of theories from the PP stable on offer.

Coherent Infomax describes the internal logic of the consumer, but for this internal logic to result in genuine representation of the external world, it must be grounded in a theory of environmental structure. Why does maximizing coherence lead to truth rather than mere hallucination? To answer this, we must connect the biological imperative of efficiency with the statistical structure of reality itself, turning to rate-distortion theory to explain how compression finds the 'natural joints' of the world.

6.3 Grounding Content: Representations as Rate-Distortion Sweet Spots

To ground the biological mechanism of coherent infomax in environmental structure, we turn to rate-distortion theory as developed by Manolo Martinez (2019a). Rate-distortion theory constitutes a branch of information theory concerned with the trade-offs inherent in lossy compression. It defines the relationship between the rate (R), or the number of bits used to encode a signal,

and the distortion (D), or the loss of fidelity in the reconstruction of that signal. In biological systems, where metabolic bandwidth is strictly limited, communication is inevitably lossy. For example, a vervet monkey possesses a finite repertoire of alarm calls to encode an infinite variety of predator scenarios. The fundamental problem for the organism is to compress this infinite variance into a finite code without losing the information critical for survival.

Martinez argues that genuine representations emerge at specific points on the rate-distortion curve known as sweet spots. These are inflection points where the curve creates a "knee"—a region where a small investment in additional bandwidth yields a massive reduction in distortion, but beyond which further increases in rate yield diminishing returns. A signal qualifies as a representation if it satisfies two conditions. First, the existence condition requires that the environment itself possess a statistical structure that creates these clumps of probability, allowing for efficient compression. Second, the optimality condition requires that the organism's encoding strategy occupy the vicinity of such a sweet spot.

This framework provides a rigorous solution to the problem of semantic smuggling identified in our critique of Miłkowski. The question of how a cognitive system knows which information sources to compare or group is resolved not by a priori semantic knowledge, but by the statistical structure of the environment. The world is not a uniform distribution of independent variables; it is populated by 'homeostatic property clusters' or natural kinds. These clusters create massive redundancies in the sensory data—for instance, the visual features of a leopard (spots, motion, shape) statistically covary with high reliability.

This statistical clustering creates a phenomenon Martinez describes as reference magnetism. Any efficient compressor, seeking to minimize distortion under a bandwidth constraint, will mathematically gravitate toward these natural joints in

the world because they are the most economical features to encode. The system does not need to know what a "leopard" is to represent it; it simply needs to follow the gradient of the rate-distortion curve to the sweet spot where the cluster is located. Thus, the "common cause" in the world creates the channel conditions that validate the comparison between different sensory streams. The semantic content of the representation is grounded in these source statistics, replacing the need for a homuncular interpreter with the blind efficacy of optimal compression.

Importantly, Phillips' theory of coherent infomax can be understood as the biological implementation of this rate-distortion strategy[55]. The partial information decomposition (PID) of the neural signal reveals how the two-point processor executes this compression. The unique information (Unq) corresponds to the uncorrelated variance or noise that lies outside the sweet spot; this is the distortion that the system accepts to maintain efficiency. The shared source information (Shd) corresponds to the redundancy created by the homeostatic property cluster; this is the signal that defines the sweet spot.

The mechanism of apical amplification acts as the physical operator that locates this sweet spot. By amplifying the neural output only when the basal input (content) coincides with the apical input (context), the neuron filters out the unique noise and propagates the shared structure. This results in a "many-to-one" mapping akin to Sterelny's concept of robust tracking, where a wide variety of specific sensory instantiations (e.g., a leopard seen from different angles) trigger the same cooperative burst state. The burst code is the compressed symbol for the natural kind, discarding the pixel-level variations that do not contribute to the shared structure.

---

[55] To clarify - Phillips is not arguing for utilizing the RDT toolbox in his account, nor he mentions it in his book. It is a proposed extension of his framework that I am advocating for on the basis of his results and Martinez's (2019a) oft-quoted here paper.

6.4 Conclusion: Semantics Without Magic

This synthesis demonstrates that there is no "magic semantics" required to bridge the gap between information processing and representation. Semantic content is the artifact of optimal compression applied to a structured world. Phillips provides the micro-level biophysical how—the separation of driving and modulating streams and the burst code—while Martinez provides the macro-level theoretical why—the necessity of finding rate-distortion sweet spots in a bandwidth-constrained environment.

This synthesis embodies what Phillips terms 'cellular psychology'—an approach analogous to cellular biology's shift of explanatory focus. Just as cellular biology demonstrated that foundational biological processes (metabolism, signal transduction, gene expression) occur at the cellular level without claiming that all biological phenomena (ecosystems, mating patterns, behavioral dynamics) reduce to cellular mechanisms, cellular psychology demonstrates that foundational cognitive operations (coherence detection, signal validation, context-sensitive amplification) are implemented at the level of individual neurons without claiming that higher-order cognition (reasoning, planning, consciousness) is exhaustively explained by single-cell properties. The neuron is not a homunculus that "decides" or "knows"; it is a biophysical device whose architectural constraints and connectivity patterns implement specific information-theoretic operations. What cellular psychology vindicates is the claim that these operations—disambiguation, grouping, precision-weighting—have a principled explanation at the cellular level, even as they participate in network-level dynamics that give rise to genuinely psychological phenomena. The explanatory payoff is not reductionism but proper multilevel integration: we now understand which aspects of cognition are grounded in cellular mechanisms and which emerge from their orchestration.

While apical amplification is our best current candidate for implementing this mechanism, the validity of the Neo-IPP's explanatory strategy does not stand or fall on this specific biological bet. The Neo-IPP is a framework that admits of multiple realizations; as we will see in the Final Conclusions, the path to a robust, naturalistic computationalism may well lead through other biological gates, provided they satisfy these same functional constraints.

Crucially, the logic of the compressed code discussed in this section is not an end in itself. As we will explore in subsequent sections, the efficiency of this compression is precisely what enables the system to reverse the process. The compressed vector, locked onto the invariant structure of the world, serves as the stable seed for generative processes, allowing the brain to decompress these representations into the rich simulations required for planning and counterfactual reasoning. Thus, the rigor of source coding becomes the foundation for the generativity of thought.

With the integration of rate-distortion theory, the theoretical synthesis is complemented and tentatively complete. We have moved from the biophysical reality of the two-point processor to the algorithmic logic of coherent infomax, and finally to the mathematical grounding of representational sweet spots. This provides a robust, naturalistic account of how neural systems can be content-bearing without requiring magic and anthropomorphic escapes. Having thus secured the foundations of information, representation, format, and content, we are now in a position to step back and survey the resulting theoretical landscape. This brings us to the conclusion of Part 1, where we can summarize the defense mounted for the view of the cognitive (neural) system as an information processor that is properly formatted, representational, and content-bearing, completing the trifecta of explanatory goals set by the (neo-)IPP.

Chapter 5: Conclusions of Part 1

1. The State of the Embattled Paradigm

We began this inquiry by acknowledging that the information processing paradigm is under siege. For decades, the "computer metaphor" served as the foundational axiom of cognitive science, providing a unifying language for psychology, neuroscience, and artificial intelligence.
Yet, as detailed in Chapter 1, the IPP has faced a sustained siege from enactivists, dynamicists, and critics of the neural code. Critics like Epstein, Brette, and Bickhard have variously attacked the 'storehouse' memory metaphor, the reality of the neural code, and the coherence of encodingism itself. However, as we have argued, these critiques do not necessitate abandoning the paradigm, but rather refining it into a more robust, mechanistically grounded framework.

Faced with this assault, one might be tempted to abandon the paradigm entirely. However, the rumors of the IPP's death have been greatly exaggerated. Despite the conceptual battering, the explanatory necessity of computation remains. As Peter Sterling and Simon Laughlin argue in *Principles of Neural Design* (2015), the staggering computational power of the brain is not an accident but a necessity of biological design. To pack "staggering computational power into a few kilograms, with power dissipation of just a hundred watts, evolution has had to find and exploit a lot of tricks." They argue that the demands on neural performance are so severe that "truly optimal design wins," implying that the specific information processing architectures we observe are not accidental spandrels but optimized solutions to computational problems.

Furthermore, Gallistel and King (2009) remind us in *Memory and the Computational Brain* that abandoning the computational framework leaves us without a mechanism for the most basic cognitive functions. They argue, "There must be an addressable read/write memory mechanism in brains that encodes information received by the brain into symbols (writes), locates the information when needed (addresses), and transports it to computational machinery that makes productive use of the information (reads)."[56] They posit that such a memory mechanism is "indispensable in powerful computing devices, and the behavioral data imply that brains are powerful organs of computation". Without the concepts of code, address, and retrieval, we are left with a physiology that cannot explain the coherence of behavior over time. The code concept, which had no place in the pre-computational framework, is foundational to the new one, just as the genetic code transformed biochemistry.

Let's recapitulate what we discussed so far through the lens of moves available for an IPP proponent. To keep the explanatory strategy fruitful but integrative (accommodating criticisms), three strategical explanatory moves are available.

---

[56] Even though Gallistel and King were rightfully critiqued by Morgan (2020) for defending an antiquated adherence to von Neumann's architecture as a strict model for the brain, it does not invalidate the general commitments of the IPP (with which Morgan agrees). What the computer paradigm needs is an update, not erasing.

2. Partition

Poland means partitions[57], as every schoolchild in this country has been diligently trained. Accordingly, IPP Warsaw-style is perfectly fine with conceding the low ground to the critics to secure the high ground. We admit that not all behavior requires complex internal representation. There can surely be different modules (although the separation works on a bigger spatial scale than modules; I am more inclined to agree with Krakauer (personal communication) that the actual iron curtain still remains on the System 1/System 2 type of explanations (Krakauer et al.,2017; for a dissenting voice: Keren & Schul, 2009[58]). But even in this text we saw successful strategies of complementary parts in play. As we saw in Orlandi's case, separating early vision and processes fueled by sensory representations can (and should) be distinguished from regimented operations contingent on representations of rules; we must distinguish between wired transitions and representational transitions. As we remember, the former are characterized by automatic, biased, embedded processes governed inflexibly by the causal makeup of the environment. Here, the critics (enactivists, Cisek) are correct: there is no "read

---

[57] Note to non-Polish readers: this refers to the process from the history of 18th-century Poland when the First Polish Republic was picked apart by three neighboring superpowers of the time - Prussia, Austria, and Russia; this is known as The Partitions [*Rozbiory* in Polish]. As a result, Poland ceased to exist as a sovereign state until the end of World War I. This became a label that became synonymous with Polish history for a lot of locally educated people and is still routinely motivated to explain differences between parts of current-day Poland.

[58] I acknowledge that invoking the 'System 1/System 2' distinction (or Type 1/Type 2 processing) invites a massive and contentious literature on dual-process theories (e.g., Evans & Stanovich, 2013; De Neys, 2012). While this debate is foundational for cognitive psychology, fully adjudicating the precise boundaries, interaction, or potential continuity between these systems lies beyond the scope of this project. For the present purpose of partitioning the explanatory landscape of the Neo-IPP, it suffices to treat the distinction as a heuristic for two functionally distinct regimes of information processing: one that is automatic, encapsulated, and 'wired' (the domain of reflexes and early perception), and another that is deliberative, decouplable, and representational (the domain of planning and reasoning). Whether these regimes map onto distinct neural architectures or are merely poles on a continuum does not seem to me like a first-priority problem for the defender of IPP, as the details of cartography and tightness of their boundaries (i.e. encapsulation) are both *prima facie* compatible with the commitments to cognitive information processing in general.

out" of a rule, just physical causation. The system "acts in accordance" with the rule (to use Fodor's term) but does not follow it. In these cases, the dynamic picture of continuous coupling with the environment is sufficient, and the "tools of information decomposition are at their weakest" (Clark, 1997).

This is distinct from representational transitions, however, that span flexible, context-dependent behaviors—such as planning, counterfactual reasoning, or language. Here, the system must physically "read" an internal data structure (a representation) to determine the next step. This preserves the IPP for high-level cognition. By adopting this partition, the Neo-IPP absorbs the insights of the "replacement hypothesis" (Wilson & Golonka) without surrendering the core of cognition. We accept that simple sensorimotor loops are dynamical and non-representational. But we insist that "representation-hungry" problems—those involving absence, error, and distal targets—require the computational machinery of the symbols factory, not just the flow of the electric pipeline.

3. The Explanatory Cascade (A Sample Of)

However, identifying *when* we need representations is often not enough; we must explain *how* they arise physically without succumbing to the "homunculus" fallacy or the "indeterminacy" problem. The "winning recipe" of the Neo-IPP is an explanatory *cascade* (to borrow this idea and label from: (Franks, 1995)) that links physical cause to semantic content. This cascade transforms the scientist's

correlational measure[59] (**a-information**) into the system's functional substrate (**f-information**). It also paves the way for successful methodological unification.

3.1 Physical Grounding: Minimal Causes

We begin with the physical layer. To avoid the charge that "information" is merely an observer's gloss on continuous physics, we can adopt Artiga's criterion of Minimal Causes. To count as a signal, a neural event must be the minimal sufficient physical cause of the adaptive behavior. This answers Brette's critique: quite bluntly, the signal is real because it does work. We offer a more sophisticated theory of the "why it works" in the chapter on similarity-based coding and its import on non-arbitrariness of coding (contrary to critics). But for now it suffices to say that even correlational binding (even if Mante and Fairhall are right in refining our too simplistic understanding of tuning curves and selectivity, and they most plausibly are right about that[60]) allows us to isolate the specific part of the spike train—the burst, the rate, or the phase—that is causally necessary for the downstream effect, filtering out the noise that might be informationally rich but causally inert.

3.2 Structural Format: Source Coding

Next, we analyze the signal's structure. Like Martinez (2019a) one can argue that the brain maximizes source coding efficiency—compressing the signal to fit the channel capacity while maintaining high mutual information with the environment. This

---

[59] In a recent paper, Lombardi et al. (2016) caution against the increasing tendency to conflate the "bit" as an abstract unit of measurement with the physical two-state system used to encode it, a confusion that inappropriately reifies the concept of information (treats it as a commodity or substrate). They argue that defining information via its physical carrier is a category error, analogous to confusing the abstract concept of a "meter" with the specific platinum-iridium bar historically used to represent it. To resolve this, they advocate for strictly distinguishing between the quantity of information generated by a source and the specific physical systems (which they propose calling "chits" rather than "bits") used to implement the coding alphabet.

[60] Anyway, a philosopher is hard-pressed to challenge that from an empirical standpoint.

explains the format of the representation (e.g., why V1 uses sparse coding) without yet appealing to semantic content. It aligns the scientist's measure of efficiency with the system's biological constraints. Furthermore, Martinez's appeal to Homeostatic Property Clusters (HPC) allows us to solve the input problem: the system tracks clusters of properties that hang together in the world, providing a structural basis for the stability of the signal.

3.3 Semantic Content: Unmediated Explanatory Information

Finally, we fix the content using Nicholas Shea's concept of Unmediated Explanatory Information (UE). The system correlates with many things (black dots, photons, flies). But the content is the specific distal condition (nutritious object) that plays an unmediated role in explaining why the behavior was stabilized by evolution. We integrate this with Israel and Perry's constraint-based framework: the neural signal carries information about the distal target relative to the specific constraints (natural laws) that the organism's lineage has tuned itself to exploit.

This cascade can be very crudely understood as Cause ↠ Format ↠ Content ↠ Behavior[61], following the structure of our analyses so far. I argue that such a strategy helps us avoid the Decoder's Dilemma. A scientist can decode many things (correlations), but only the information that satisfies the UE criterion for the system's specific task function counts as f-information (information-for-the-system). The "homunculus" is replaced by the consumer system—the downstream mechanism evolved to exploit this specific informational format. The explanatory cascade provides the naturalistic criteria for identifying exactly where this reading and writing occurs.

---

[61] We have not touched upon this transition, but it is precisely the point of Ritchie and colleagues solution to the problem they state. In short, according to them, MVPA-type decodings have to be tied to psychological models, which are thus explanatory of content being causally responsible for behavior (and not only reliably but potentially spuriously correlated with).

4. Methodological Pluralism

The final pillar of the Neo-IPP is the rejection of the zero-sum game between computation and dynamics. We advocate for methodological pluralism, viewing IPP and dynamical systems theory (DST) not as competing ontologies but as complementary descriptions of the same mechanism.

Contrary to radical critics of information-based paradigms, the IPP does not require information processing to be the sole component in a successful explanatory strategy. What is at stake is the productiveness and necessity of IPP core tenets and posits. Surely, information processing is not sufficient for explaining cognition, but equally surely it is necessary.

As Massaro (1990) noted, "Most environment-behavior relationships are too complex to be predicted from simple information processing principles... This failure does not reflect the limitations of the principles, as much as the complexity of the natural world" (p. 162).

Point taken. Luckily, IPP as an explanatory strategy is not militant towards most other strategies in the philosophy of explanations in cognitive science. Particularly causal and neomechanistic strategies, which I intend to argue for in this thesis. To merely foretell some of the conclusions, I claim that the resolution of this conflict lies not in choosing between "mechanism" and "information," but in recognizing their deep complementarity. Writings of the "New Mechanists", such as Bechtel, Machamer, Craver, and Darden, have redefined "mechanism" in a way that explicitly accommodates and even requires the IPP (Machamer et al.,2000; but even more: Bechtel, 2008). *Contra* Brette's concern that "coding" ignores causality, William Bechtel argues that "information processing" is simply a characterization of the operations within a mechanism (Bechtel, 2011). A mechanism is defined as "a structure performing a function in virtue of its component parts, component

operations, and their organization" (Bechtel & Abrahamsen, 2005: 423). In cognitive mechanisms, the "parts" are neural representations, and the "operations" are computations. This view does not require the brain to be a linear machine; as Craver and Darden note (2013: 20), mechanisms can be distributed, nonlinear, and active. Bechtel emphasizes that characterizing operations as information processing allows scientists to explain how mechanisms regulate themselves and adapt to environmental variability—a crucial biological function that purely energetic descriptions miss.

From the causal side of things, Lauren Ross offers the solution to erase the gap further by clarifying the nature of causal explanation in neuroscience (Ross & Bassett, 2024). She argues that "causal information" provides explanatory force by identifying the factors that control an outcome. Specifically, she suggests that pathways in the brain can be understood as causal constraints—structures that limit and direct the flow of energy and influence. The IPP, in this view, describes the specific topography of these constraints. When we say a circuit "processes information," we are identifying the reliable causal pathways that allow specific inputs (e.g., retinal patterns) to control specific outputs (e.g., motor commands).

This aligns with a host of explanatory proposals already firmly in place in the literature, such as the already discussed Beer and Williams' (2015) proposal to treat IT and DST as "different mathematical lenses." Information theory quantifies the reduction of uncertainty and the structure of dependency, while dynamical systems theory quantifies the stability and trajectory of the states. Neither description is exhaustive; both are necessary to explain how a physical system achieves cognitive success.

IPP could be cashed out as mechanistic computationalism (Miłkowski, 2013), a view that treats computation not as a semantic interpretation but as a physical process.

On this view, "mechanistic computationalism is compatible with various approaches" because it defines computation as the manipulation of vehicles according to rules, regardless of whether those vehicles are discrete symbols or continuous variables.

Hence, we can use his (Miłkowski, 2016; Miłkowski & Hohol, 2021) position on unificatory strategies in cognitive science. After all, he argues that "computationalism is only plausible if you also accept explanatory pluralism" (Miłkowski, 2013). We can describe the brain physically (as a dynamical system of ions, dendrites, membranes, nooks, and crannies) and computationally (as an information processing system) without contradiction. The computational description is valid because it picks out the functional organization of the physical mechanism. By treating information as the *function* of the neural *mechanism*, we resolve the "mongrel" accusations. To use a computationalist tale of all times, the code is not a ghost in the machine, but the organization of the machine itself.

5. The Neo-IPP Commitments

The information processing paradigm is far from the finished product, and neither is it destined for adopting one strict explanatory strategy (as we saw above while assessing different paths of life for the less-combative computationalists). Nonetheless, this does not preclude us from delineating a set of hardcore commitments that any honest-to-god information-flow psychologist (to remain strictly Fodorian or an information-flow cognition theorist more broadly) should reasonably adhere to and hold their ground.

IPP has survived its purported deconstruction not by ignoring its critics, but by evolving. It has moved from the "empty metaphor" criticized by Epstein to the

"mechanistic computationalism" described by Miłkowski (2018). By distinguishing between "wired" reflexes and "representational" thought, and by grounding "information" in rigorous causal necessity, the field has moved toward a robust, biologically grounded science of the mind.

This evolution is perhaps best illustrated by our treatment of memory. While classical proponents like Gallistel and King (2010) correctly identified the *functional necessity* of a 'read/write' mechanism for generative cognition, they tethered this insight to an outdated architectural demand—that the brain must implement a literal Turing tape. As Alex Morgan (2020) argues in *Against Neuroclassicism*, this confuses the abstract logic of computation with its specific physical implementation. The Neo-IPP accepts Morgan's correction: we rescue the indispensable *computational function* of the memory mechanism (variable binding) while rejecting the *classical architecture* in favor of neural implementations, such as the recurrent attractors and two-point processors discussed in Chapter 3.

To solidify the standing of this evolved paradigm, we can summarize the "Neo-IPP" as a research program defined by adherence to five core commitments. These tenets trace their lineage back to the foundational arguments for functionalism in Fodor's *Psychological Explanation* (1968), but are here updated to accommodate the biological constraints of the neural age.

A theory qualifies as a valid instantiation of this paradigm if and only if it endorses the following theses, core tenets of the updated IPP (thus: *neo*-IPP):

T1 The Foundational Thesis (Information Realism): Cognition is not merely accompanied by, but is constituted by, the transformation of information vehicles. Nervous systems are functional mechanisms where physical states carry structural information that is causally relevant to the system's output. This distinguishes the

cognitive system from other physical systems where information is merely an observer's gloss.

T2: The Architectural Thesis (Decomposability): The cognitive system is explainable via decomposition into functional components or "stages" (e.g., encoding, comparison, decision). While strict linearity is not required, the system must possess a functional architecture where the outputs of specific operations serve as inputs for downstream consumers. This validates the search for the "Algorithmic Homunculus"—the subsystem that consumes the output.

T3: The Mechanistic Thesis (Strong Equivalence): Computational descriptions are not merely instrumental or metaphorical; they must map onto the biological reality of the neural substrate. A valid IPP model must demonstrate how algorithms are implemented by neural dynamics, moving from "weak equivalence" (input-output matching) to "strong equivalence" (process matching).

T5: The Representational Thesis (Content Specificity) For high-level, flexible processing (Decoupled Systems), the system must utilize internal data structures (representations) that stand in for absent or abstract environmental features. These representations are distinct from automatic, "wired" physical couplings and are defined by their specific semantic content (grounded in Homeostatic Property Clusters and Teleosemantics) and physical format (grounded in source coding and computational accessibility).

T6: The Decodability Thesis (Causal Readout): The information postulated by the theory must be empirically recoverable from the physical substrate (e.g., via neural decoding). However, successful decoding is subject to the f-information constraint: the information counted as "cognitive" is not merely the statistical correlation available to the scientist (a-information), but the information that is causally exploited by a downstream mechanism to guide adaptive behavior.

6. Conclusion: The Only Game in Town (Again)

Putting it a bit pompously, we ascend in science not by discarding our best tools when they reveal complexities, but by refining them. That includes the conceptual tools for thinking that are theories and frameworks. The critique that the brain "is not a computer" because it does not resemble a laptop often rests on a category mismatch. It risks confusing the artifact (the silicon chip) with the principle (computation). As Shagrir asks in the opening of *The Nature of Physical Computation*:

> "What is it about laptops, smartphones, and nervous systems that they are deemed to compute—and why does it seldom occur to us to describe stomachs, hurricanes, rocks, or chairs that way?" (2022:1).

The answer is that only the former systems possess the functional architecture to manipulate vehicles in virtue of their content. Stomachs digest; hurricanes blow; but neither processes information to solve a problem that is decoupled from its immediate physical instantiation. The IPP remains "the only game in town" because it is the only framework capable of explaining how the semantic properties of symbols - when properly formatted and contentfully constrained - can be causally efficacious in a world of physical forces. Without it, we are left with a dualism of "causes" (neurobiology) and "reasons" (psychology) with no bridge to connect them.

The critics who demand a perfect, gapless correspondence between our computational models and the biological wetware fall prey to the trap famously described by Borges in *On Exactitude in Science*. He imagines an empire where the Art of Cartography attained such perfection that "the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it." The following generations, realizing that such a map was useless, "delivered it up to the Inclemencies of Sun and Winters".

The 4E critics are wrong in accusing IPP of an overintellectualized picture that demands a theory for the Map of the Empire. At the same time, but on the other side of the spectrum, resting solely on dynamic coupling of the body and environment amounts to sailing through tricky waters in changing conditions only knowing a few rigid tricks and heuristics. That only gets you so far, and more often than not, that so-far is the seabed.

To resurface, we might borrow a famous quip from Edsger Dijkstra, who claimed that asking whether a computer "thinks" is no more interesting than asking "whether a submarine can swim." Dijkstra dismissed this as a pseudoproblem, linguistic trap, but his metaphor inadvertently fits our point. After all, trying to navigate the depths by opening the windows to "let the environment in" is rarely a successful strategy.

The Neo-IPP offers something far more useful: a map that abstracts, simplifies, and formats the territory into a structure we can understand and utilize effectively. It is the source of the awe-inspiring quickness, flexibility, and accuracy of higher-order cognition. It provides the fundamental concept on which the entire interface of theoretical posits rests, *via* which we can grasp the mind. By integrating the rigorous causal constraints of modern neuroscience with the functional architecture of psychology, the Neo-IPP proves that while the map is not the territory, it is the only way we can navigate it.

Onwards, Beyond Blobology

This commitment to decodability finds its empirical manifesto in Russ Poldrack's *The New Mind Readers* (2018). In the book, he diagnoses the field's historical struggle with "reverse inference"—the difficulty of deducing cognitive function from neural activation—not as a fatal flaw of the computational view, but as a failure of methodology. He argues that for too long, neuroscience has been stuck in the era of

"blobology": the search for *where* activation occurs (localization), rather than *what* is being processed (content). Blobology (as a sort of new phrenology indeed; Uttal, 2003) yielded only correlations—statistical maps of blood flow that remain vulnerable to Nizami's critique of being observer-relative artifacts.

However, Poldrack demonstrates that the field (and subsequently, cognitive neuroscience as an ensemble) is undergoing a paradigm shift from "where" to "what"—from localization to decoding. This shift allows us to defeat the challenge posed by critics like Nizami, Brette, and Bickhard, who argue that neural "codes" are merely metaphors or observer projections. Poldrack argues that by using Multivoxel Pattern Analysis (MVPA; Haxby et al.,1991) and Representational Similarity Analysis (RSA; Kriegeskorte et al.,2008; Kriegeskorte & Kievit, 2013; see: Roskies, 2021 for a philosophical discussion), we can distinguish between mere correlation (between the stimulus and the blob) and genuine computation (the processing of causally efficacious information that generates behavior). While correlations can be spurious, computation leaves a structural signature: a representational geometry that is preserved across different individuals and different tasks. We will analyze this point thoroughly while offering a novel way of tying RSA to the logic of information in chapter 5.

For now, it suffices to notice that, crucially, Poldrack's project relies on the very premise that the Neo-IPP defends: that neural patterns are not merely metabolic shadows, but structured formats that carry specific content. As Poldrack notes, our ability to decode mental content is limited only by the granularity of our computational models. This vindicates the Neo-IPP strategy: we solve the 'Decoder's Dilemma' not by retreating from computation, but - to use the distinction we coined here - by ensuring our decoding algorithms track the f-information (functional content) defined by the system's own representational architecture, rather than the mere a-information (statistical artifacts) of the

observer. In this sense, the 'New Mind Readers' are simply the rigorous cartographers of the Neo-Computationalist map.

This success provides a transcendental argument for the IPP: explanatory mindreading is feasible *if and only if* the information processing paradigm is true. If the enactivists were correct—if the brain were merely a dynamical system coupled to the world without internal, content-bearing vehicles—then there would be nothing to decode. The statistical patterns of MVPA would be empty correlations, incapable of predicting novel mental states. The fact that we *can* decode the specific content of a thought (e.g., "imagining playing tennis") from the structural format of neural activity proves that the brain is not just dynamically active; it is trafficking in formatted, contentful vehicles. Thus, Poldrack's "new mind readers" are not imposing a computer metaphor on the brain; they are discovering the f-information that the brain itself uses to solve problems.

PART 2: APPLICATIONS AND EXTENSIONS

From Construction to Stress-Testing

In Part 1, we constructed a defensive perimeter around the Information Processing Paradigm (IPP). Facing the siege of enactivism, interactivism, dynamicism, and eliminativism, we argued that the paradigm does not need to be abandoned but rather reconstructed. The result was the Neo-IPP: a framework defined by specific commitments to structural formats, unmediated explanatory information (UE), and mechanistic computationalism (among others). I concluded that if the cognitive system is partitioned correctly—distinguishing "wired" sensorimotor loops from "representational" decoupled reasoning—the computational stance remains the "only game in town" for explaining high-level cognition. It was also argued that defending an IPP-type strategy could be reconciled with causal and interventionist explanations

However, a theoretical defense is only as strong as its explanatory purchase. A paradigm is vindicated not by its logical coherence alone, but by its ability to solve specific, intractable problems in the special sciences. If the Neo-IPP is valid, it must do work. It must clarify the opaque operations of artificial intelligence; it must resolve the ancient tension between causal mechanism and logical computation; and it must explain the generative capacities of the mind.

Part 2 is dedicated to this task. It comprises three standalone investigations, each designed to stress-test a specific pillar of the Neo-IPP framework against a concrete domain of inquiry. These chapters move from the synthetic (artificial intelligence) to the philosophical (philosophy of science) and finally to the generative (cognitive modeling).

The Logic of the Extension

Chapter 6 (Representing as Compressing) tests the representational thesis (content specificity) and the concept of format. In Part 1, we argued that representation

requires specific formats (e.g., polar vs. Cartesian) that render information computationally accessible. Chapter 6 applies this to the "black box" problem in deep learning and perceptual psychology. It argues that deep neural networks (DNNs) are engines of transformational abstraction—mechanisms designed to discover optimized formats that disentangle the causal structure of the world. This chapter vindicates the empiricist commitment of the Neo-IPP: that structure can be learned from statistics, provided the architecture is biased toward the right kind of format (compression).

Chapter 7 (Meaning from Matching) tests the ‘stability’ of informational structures and their causal activity. A central anxiety of the siege was the claim that computational explanations are merely "intentional glosses" that ignore the causal reality of the brain. Chapter 7 attacks this problem head-on by defending a Correspondence Network Framework. It demonstrates that computational explanation is not a rival to causal-mechanistic explanation but a subtype of it. The crucial move defended in this chapter is marrying Correspondence Theory of Semantic Information by Miłkowski to Representational Similarity Analysis (RSA) in neuroscience, arguing the former provides a theoretical base and justification for why certain representational geometries are stable and mutually informative (as discovered by the RSA). By analyzing RSA as a mechanism for detecting cross-channel consistency, this chapter provides the rigorous philosophical proof that structural similarity is not just a metaphor but a causal mechanism for grounding content.

Chapter 8 (On Generating Information) tests the generative power of the Neo-IPP. In Part 1, we defended the structural format thesis: the idea that cognitive systems must compress environmental information into efficient, disentangled formats tokened according to the needs of the consumers. But efficiency is not an end in itself. Why does the brain invest metabolic cost in building these complex,

compressed models? Chapter 8 provides the answer: generativity. It argues that the ability to generate novel, plausible information (extrapolation)—rather than merely retrieving stored data (interpolation)—depends entirely on the quality of the underlying compression. By analyzing case studies in computational photography, this chapter demonstrates that "genuine generativity" cannot be accounted for solely based on engineering principles, but neither do existing philosophical frameworks for obtaining new knowledge apply. In turn, a new cluster definition of genuine generativity is offered that combines compression, internal kernels for new iterations, and causal rule-following (Taking condition).

Together, these chapters demonstrate that the Neo-IPP is not merely a defensive bunker for an old theory but a generative framework for the future of cognitive science.

Chapter 6: Representing as Compressing

Introduction: Why compression is imperative

In Part 1, we established the core commitments of the Neo-Information Processing Paradigm. We argued that for a cognitive system to be representational, it must traffic in vehicles that are structurally formatted, decouplable, and semantically grounded. However, a theoretical defense is only as strong as its explanatory purchase. We also promised that we must now "stress test" these commitments against the rigorous demands of empirical performance. The first and most fundamental of these tests concerns the economy of the mind. If the brain is indeed an information processor, it is one operating under severe metabolic and channel-capacity constraints. This chapter argues that the first virtue of a representational system is not richness, but compression. By analyzing the link between redundancy reduction and generalization, we will demonstrate that the "format" defended in Chapter 3 is fundamentally a compressed format and that representation itself can be fruitfully understood as the product of efficient coding.

The structure of this chapter is as follows: in Section 1 I discuss some quite general principles connected to compression-like processes that fit nicely with what we should reasonably expect our theories of perception to explain. In Section 2, I offer empirical evidence from a wide range of cases that demonstrates the robust presence of compression-like mechanisms in the perceptual machinery of human and artificial agents and on different levels of processing. In Section 3, I address possible worries about the fuzzy conceptual status of the notion of information, which could endanger the promise for its explanatory capacities. Then I shift to analyze the consequences of looking more deeply at the nature of perceptual

representation as information processing. Finally, I conclude in Section 4 with speculation on whether this strategy might be helpful in assessing structural similarities in terms of infomorphisms.

1. Regularities, compression, and efficient perceptual coding.

'Brains are the ultimate compression and communication systems'—David MacKay declares at the beginning of his monumental *Information Theory, Inference, and Learning Algorithms* (MacKay, 2003:V), a foundational book for computational modeling of cognitive processes. Chaitin (2006) thinks that 'compression is comprehension.' Similarly, Schmidhuber (2009) developed the idea of data compression as a basis for prediction in cognition. Compression-based explanations are found in neuroscience (e.g., Allen-Zhu et al.,2014), psychology (Wolff, 1993), and recently, in machine learning (Goodfellow et al.,2016). It seems plausible to assume that there might be a philosophically interesting connection between these quite diverse applications and perhaps some wiggle room for identifying structural similarities in the virtue of convergent information processing.

Although quite a rare occurrence, philosophers do sometimes express interest in compression. For example, recently Daniel Wilkenfeld (2019) put forward a theory of 'understanding as compression.' He articulates this idea as follows:

> *'Understanding as Compression (UC):*
> A person p1 understands object o in context C more than another person p2 in C to the extent that p1 has a representation/process pair that can generate more information of a kind that is useful in C about o (including at least some higher order information about which information is relevant in C)

from an accurate more minimal description length.' (Wilkenfeld, 2019:5)

Obviously this definition pertains to a high-order, personal level of cognition, as one would reasonably expect. Wilkenfeld is quick to admit that, as he states, he is concerned with what understanding consists of for persons and concedes that might not be directly applicable to the types of compression we are most used to—artificial algorithms that usually deal with arbitrarily long strings of digits (or other mathematical representations) that are usually intractable for humans. Hence, it might be surprising that there is much to gain by tying compressing processes with human perceptual powers. But the general idea that motivates Wilkenfeld's account is a significant one.

He claims that both compression and understanding share their generative nature. To take his point across, he gives an example of two people—a novice and an expert trying to prove a particular theorem—a soundness proof for first-order systems. A novice uses the most mechanical memorization strategy (with the Sheffer stroke operator [NAND], as it is a universal gate[62]), whereas the expert uses generalized knowledge of the matter to generate a full proof from simpler internal representations. Wilkenfeld notes that if any of the two provers happen to miss a step, they find themselves in wildly different situations. The novice faces a daunting predicament, as he is unable to notice and fill in the gap. The expert, however, can comfortably do that by generating missing pieces of information from simpler representations. By invoking his definition, the expert understands better as she can generate more contextually relevant information from minimal description (generalized knowledge of the problems contrary to the full-blown memorization). The invoked idea of compressing up to the 'minimal description length' is directly

[62] I.e., all three elements of the functionally complete set {AND, OR, NOT} can be construed using only [NAND]

borrowed from Grünwald (2007), which neatly captures the intuitive idea that 'the more we can compress the data, the more we have learned about the data' and directly nods to the algorithmic interpretation of information theory (Chaitin, 1977).

Wilkenfeld's story demonstrates at least two of the properties that qualify compression-like processes as promising exploratory models for a host of cognitive tasks that we reasonably expect to require adherence to the same goals—generalization and generativity (Buckner, 2018). According to Wilkenfeld, 'being in such a state where you have a representational kernel and an ability to piece together the rest is not just a result of understanding, but what it means to understand' (Wilkenfeld, 2019:2). However, this idea can be understood more generally than just the feature of highest-level cognition (as understanding *prima facie* is) as pertaining to the cognitive ability of generalization (compressing[63]) and subsequent generativity (decompressing). To borrow his term, we can say that the idea of compression-like learning can be construed as a maximization of 'output-to-scaffolding ratio' (think again of the seasoned logician example, which outperforms the meager memorizer in the virtue of having been able to produce any potentially missing step).

The idea of possessing an adequate representational kernel strikes us as similar to the semantic pointer theory (Eliasmith, 2003), which also explicitly refers to the notion of compression as an important feature of cognitive information processing. Blouw et al.,(2015), for instance, define semantic pointers as 'symbol-like representations that result from the compression and recursive binding of perceptual, lexical, and motor representations, effectively integrating traditional connectionist and symbolic approaches.'

---

[63] Here in a more restricted sense—as a part of the entire procedure of compressing/decompressing pairs that form the entire schema of 'compression (in a broader sense).

Semantic pointers (Eliasmith, 2011) formalize this generalizing-generative function through the mechanism of compression via recursive binding. In this framework, high-dimensional sensory information is compressed into lower-dimensional vectors (pointers) that retain deep structural relationships, effectively functioning as symbols that can be manipulated syntactically. Crucially, this compression is reversible: the system can 'dereference' the pointer to reconstruct the original detailed sensory state.

Thus, semantic pointers are generalizing in the sense that they provide a stable, low-dimensional handle for high-level cognition while remaining generative by preserving the causal link to the sensory populations required to reconstruct the specific experience. They are generalizing in the sense that they i) point to the perceptual information stored at the lower level (say, the images of flat surfaces with legs, the concatenation of letters 't-a-b-l-e,' tactile sensations of tables, etc.); ii) retain semantic information extracted on the lower level. The generativity is supposed to come in the form of their readiness to be used to reuse the learned, compressed representation without the need to activate the perceptual-level states that the pointer 'points to.' This figure summarizes the idea with a toy example:

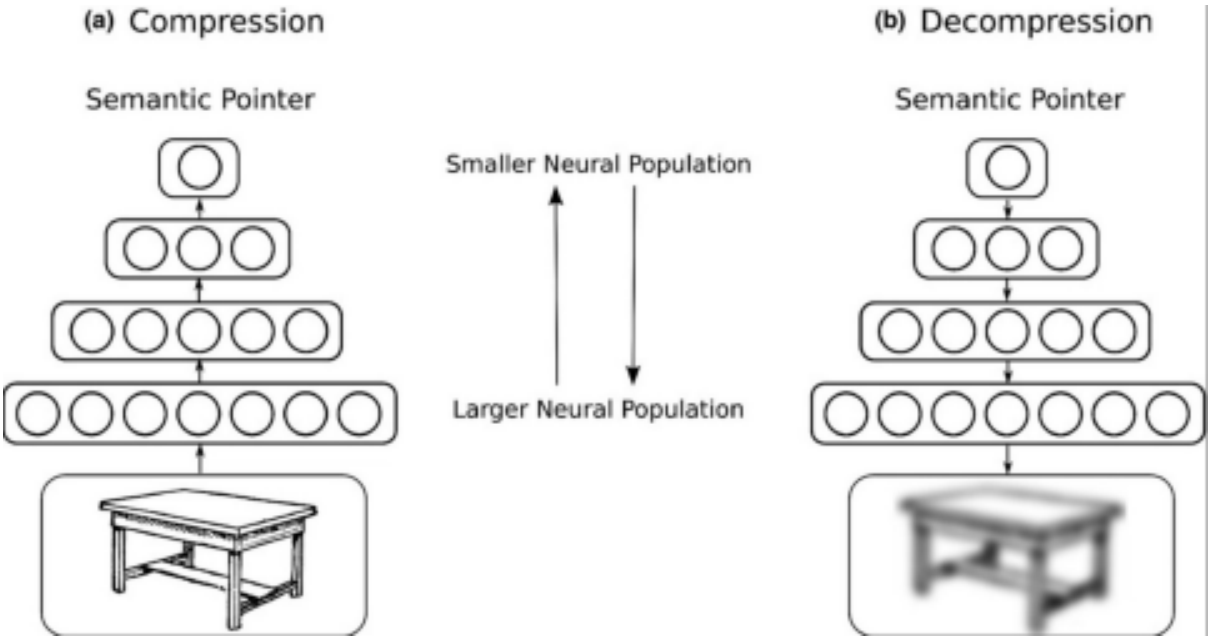

Figure 9: The schematic of the semantic pointer of TABLE. Multimodal information extracted and processed by the larger neuronal population is gradually compressed to form a more minimal representation. The subsequent evocation of the representation for some other task (e.g., recall) is achieved by the somewhat reversible procedure (compression is lossy, hence the decrease of the resolution'; adapted from Blouw et al.,2015: 1133)

A similar idea of looking for learned, *compressed* representations in information processing perceptual tasks is pursued in the self-supervised learning subfield within machine learning. For example, Hjelm et al. (2019) suggest that the technique called Deep InfoMax [DIM] achieves compressed representations by the maximization of mutual information between the encoder's input and output. This talk of representation learning and representation compression is commonplace in the subfield[64] (Poldrack, 2021). Other information-theoretic measures, such as

[64] An important caveat is in place here: both the data and the model can be compressed. The former can be conceptualized as the goal of processing; the latter is used to speed up the inference time (Goodfellow et al.,2016:448)

Kullback-Leibler divergence, are also frequently used to explain the performance of the given network.

We see here that, contrary to the semantic pointer framework (and corresponding SPAUN architecture), which invokes the semantic interpretation of 'information,' machine learning techniques directly refer to the classical sense of information. This apparent lack of clear definition of the term 'information' is a potential philosophical problem, because it might imply that there is a latent equivocation in place.

For now, it ought to suffice to state that compression is usually defined as an information processing procedure in the classical sense of the term (Shannonian) and described with tools and concepts from information theory or by its extension, which is Kolmogorov-Solomonoff-Chaitin's algorithmic interpretation. The simplest definition states that it is a process that encodes the original piece of information (signal) with the shorter string (lower bit rate) than the original formulation. This kind of operation can only succeed when there is an element of statistical redundancy in the signal to be encoded. Redundancy can appear only if some kind of regularity organizes the given signal or dataset. Entirely random processes are then fundamentally uncompressible. This crucial relation served Chaitin (1977) to define the very notion of randomness in terms of the compressibility of its respective signal. Put plainly, if there is no possibility of representing the original message in a shorter form (or, in other words, generating a shorter-than-original program code), then it is entirely random (Dennett, 1991).

Compression comes in either lossless or lossy form (Cover & Thomas, 2006). The main property of lossless compression is its reversibility. The product of such compression, i.e., transformed (shortened) version of the original message, can be

inversely brought back to the original form without any piece of information lost in the process. Lossy compression, on the other hand, is a trade-off procedure. Some information is lost in the transformation, so the process is not fully reversible, and the relata of the transformation are not mathematically identical. But the payoff in terms of the saved storage (shortened length) is usually much greater than in the lossless-type procedures (hence it's a trade-off). Hence, it is most likely to think that the human perceptual system is mostly concerned with lossy compression. Bates and Jacobs (2020), for example, claim that perceptual processes instantiate coding-procedures defined formally by the 'rate-distortion theory,' which is concerned with efficient communication with the minimization of loss of information during the encoding/decoding procedures.

But how does this elegant, yet abstract, mathematical set of ideas correspond to perceptual phenomena? The most trivial observation is that both deal with selective, regularized, generalizing information processing, so there is *some* correspondence, but that's surely not sufficient. More importantly, though, perception would not succeed in a random environment. The goal to explain perception is to explain how it operates in a highly structured, regularized environment. This regularization points at the notion of predictability. As we saw earlier, the predictability of both the stimuli and the signal is of crucial importance for efficient use of regularities underlying the informational landscape the agent operates in. That prima facie suggests that compression-like usage of the regularities can help explain the basis of the important feature of perceptual systems that is their ability to 'fill in' the details, as in numerous examples of perceptual completion in psychology (Kanizsa, 1985; Sekuler & Palmer, 1999). It can also serve as a rival model for furthering the debate on the use of purported 'knowledge' in perceptual processes. The relation I have in mind can be spelled out in a simple maxim that links perception and compression on a theoretical level:

Redundancy = regularity = predictability.

I will build on this rather general set of observations and try to analyze how they contribute to understanding perceptual processes as efficient coding procedures that are compression-like. For the sake of brevity, I restrict the scope of interest to the visual tract.

Perception is a heavily resource-limited enterprise and takes advantage of that principle. Sensory detectors are attuned only to particular classes of signal. Their resolution varies greatly within the system, too. For example, visual acuity is much greater within the fovea (where the density of rods and cones is much higher than in the peripheries), rendering the on-center (of fixation) sight much sharper than off-center (Tootell et al.,1988). But that's just the start, as just a step later the visual processing encounters the first obstacle in the form of the bandwidth of the optic nerve that relays information to the cortical structures (or rather, more precisely, first to the lateral geniculate nuclei in the thalamus that is only later imported up to the cortex). The throughput that the optic nerve can transmit is much smaller than the capabilities of the population of photoreceptors. Their output must by then be compressed in some way. The process that is responsible for that is called whitening (Field, 1987) and is performed in retinal ganglion cells, just before the signal is sent downstream to the aforementioned choke point (informational bottleneck; for the computational discussion of the idea, see Tishby and Zaslavsky, 2015). High-spatial frequencies (less common in the environment, responsible for coding abrupt changes, i.e., edges) are weighted higher than low-spatial frequencies (coding smoother transitions in the images), receive less weight and are thus effectively filtered out.

At the cortical level, this anatomical limitation is virtually absent. Neurons in the neocortex are highly interconnected, giving more degrees of freedom to implement

different ways of efficient coding. Sparse coding is one of them. In fact, different representations at the neural population could all decode the retinal input (Foldiak & Young, 1988). From the very extreme of a dense code, with only a few neurons, all being relatively active at any given input, to a very local code with many neurons but only one being active at any given input. Sparse coding is in the middle of these two extremes. But, depending on the definition of efficiency, all three codes can be considered efficient (depending on the combined cost of having many neurons, the cost of having many interconnections between neurons, metabolic firing cost[65], etc.).

Whitening is not a necessary component of sparse coding. These are two different efficient coding principles and compression procedures that seem to co-exist at different processing stages in the human brain. A few side notes here. It may well be possible that these two coding strategies are formally or mathematically related, and that both can be reduced to the same learning algorithm. This search for a single underlying principle is an active topic of research for theoretical neuroscientists. Recent advances in deep nets, or convolutional neural networks, show that some architectures automatically develop a sparse code, without a sparseness constraint being explicitly built-in (Li et al.,2013). In computer vision, pre-whitening of the input leads to faster learning of a sparse basis set of

---

[65] The energetic cost is a truly important one for the biological systems. Brains are fantastically efficient in that respect, utilizing ~20 W (which still constitutes a whopping 20% of the organism's total energetic intake) (Attwell & Laughlin, 2001). This is vastly more efficient than artificial systems; for comparison, the training of a single large language model like GPT-3 is estimated to consume 1,287 megawatt-hours (MWh) (Patterson et al., 2021), while even running such models requires industrial-scale energy that dwarfs the biological budget (Sejnowski, 2018). This disparity largely stems from the prevalence of 'Red AI'—a paradigm where time and memory are optimized to buy accuracy, while energy is treated as free real estate (Schwartz et al., 2020). This is important to bear in mind while tempted to assume that artificial systems can be optimized according to the same constraints on purely computational terms, without shifting towards the efficiency-first mindset of 'Green AI' (Schwartz et al., 2020; Strubell et al., 2019)

components (and also a less noisy basis set, Le et al.,2011). So, although whitening is not a necessary component of sparse coding, it seems to improve the results of a sparse coding[66].

However, to understand why these biological mechanisms constitute a form of representation rather than mere metabolic housekeeping, we must look to the formal constraints of efficient coding.

As we detailed in Chapter 4 regarding semantic content, rate-distortion theory (RDT) provides the formal framework for this optimization. While Martinez applied RDT to semantics, Bates and Jacobs (2020) demonstrate its necessity for perception. Why? Because perceptual systems also face a fundamental optimization problem that defines the very nature of encoding: they must minimize the bit rate (R) of the internal signal while maintaining a specific level of fidelity, or "distortion" (D), relative to the environmental source. This is not merely an engineering constraint; it is a theory of resource allocation that entails three constitutive principles of cognitive architecture.

First is the limited capacity principle. Biological channels—whether the optic nerve, the thalamocortical loop, or working memory—have finite bandwidth. This forces the system to adopt a "lossy" strategy. Unlike a digital archive that seeks perfect preservation, a cognitive system compresses data to fit the channel. Second is the prior knowledge principle. Efficient compression is mathematically impossible without a model of the source statistics. To compress a signal effectively, the brain must already possess a probability distribution (a "prior") of the environment (e.g., natural scene statistics). Thus, the very act of efficient compression implies the

[66] It might be worth pointing out that we are then involved with the algorithmic level, speaking in Marrian terms (1982; Shagrir & Bechtel, 2015).

existence of internalized structural knowledge—the system must "know" the world to code it cheaply.

Third, and most critical for the Neo-IPP, is the Task-Dependency Principle. In classical engineering, "distortion" is often measured by a generic metric like mean squared error. In biological cognition, however, the distortion function is determined by the organism's behavioral goals. A "good" compression for a predator is one that discards color information to preserve high-temporal-resolution motion cues; for a forager, the weighting is reversed. This transforms compression from a syntactic operation into a semantic one.

This aligns seamlessly with the information bottleneck (IB) framework (Tishby et al., 2000). The IB principle posits that the goal of a deep neural network (or a brain) is to squeeze the high-dimensional input X through a bottleneck variable T to predict an output Y. The optimization objective is dual: maximize the mutual information between the compressed representation and the target (I(T;Y)) while simultaneously minimizing the information kept about the raw input (I(X;T)). Under this view, "abstraction" is the mathematical result of this squeezing process. The system does not "create" abstract concepts ex nihilo; it reveals them by systematically forgetting the nuisance variables (noise) that do not contribute to the prediction of Y. Thus, lossy compression is the engine of abstraction: we lose the pixel-level details precisely to gain the causal invariants.

We will come back to the point about the striking similarities between the findings in computational neuroscience and in engineering artificial neural nets. These kinds of principled, information-based relationships suggest that the similarities are non-coincidental. That suggests that understanding artificial nets as models can truly help us to further our understanding of perceptual processes.

Please note that we can differentiate between plain insensitivity to the particular parts of the environmental signal and actual filtering. Human eyes are sensitive only to a relatively narrow spectrum of the electromagnetic spectrum (the 'visible spectrum' or the 'visual window'). We are not inclined to say that we filter out, for instance, infrared light in any sense, but rather are not equipped to handle this particular type of signal. Conversely, low-spatial frequencies (within the visual window) are actively suppressed. Certainly, the neural signal-processing picture is much more complicated and includes also non-linear procedures such as threshold and saturate functions, pooling, normalization (DiCarlo, 2012, Carandini et al.,2005, Lennie & Movshon, 2005), etc., but for our present purposes, it suffices to say that there is an important difference between certain 'deafness' (or insensitivity) to particular (potentially informative, hence behaviorally advantageous, after all) signals and their active filtering. Only the latter can be meaningfully described as compression-like, and only the latter is a perceptual activity, as is caused by the stimuli actually impinging on sensory apparatus.

We can think about this problem from a different perspective when we take into account how important the format of informational intake is. Seeing a particular image, say, a portrait, usually does not pose too great of a challenge to the human perceiver. Let's suppose, for the sake of clarity, the image is represented in grayscale of low granularity (so there are just a handful of particular shades of gray, say 10, coded from 0 to 9). The picture is still effortlessly processable and perhaps even recognizable (if the face is familiar). Now, if we 'translate' the grayscale values to their corresponding numerical values displayed as actual symbols (so every pixel on the matrix would be substituted by its respective grayscale value; e.g., all the black pixels would be substituted by zeroes, white pixels with 9s, and all the intermediate levels with respective digits). Even though that representation is still mathematically equivalent (structure-preserving, homomorphic), it ceases to be

readily available (Hyvaerinen, Hoyer, Hurri, 2009:5; we saw this example in Figures 5 & 6 in section 3.2, ch. 3).

A similar idea was discussed by Dennett (1991:33-34) when he considered translating the 900-pixel bitmap into 4-bit chunks and then substituting them with equivalent hexadecimal encoding. After such a transformation, the previously easily recognizable patterns in the bitmap (and the potential intrusion of noise) become intractable in the hexadecimal string.

We can see how even simple static-image processing can be potentially problematic for perception. But real-world visual perception is, perhaps unsurprisingly, an infinitely more complicated task. If we think of the environment as the dataset, it is as high-dimensional a set as they come. Thus, any cognitive agent that challenges it—be it a paramecium, a robot, an animal, or a human being—faces the same reality, under the enchantment of what is sometimes called in computer science 'the curse of dimensionality' (Bellman, 1956). It can be boiled down to the idea that as the number of dimensions grows, the number of samples needed to theoretically cover the entire space of possible inputs grows exponentially[67]. Any successful cognitive system has to utilize representations, because the independent 'one-to-one' extraction of every single feature of the environment in isolation is computationally impossible (Poldrack, 2021).

It really seems to be a daunting challenge. Nevertheless, somehow all the organisms that inhabit the 'world out there' face this very curse yet do not seem to care, and cope just fine. But what is perhaps even more interesting—recently the machines started to escape the curse, too. Hinton and Salakhutdinov (2006), in a classic paper,

[67] It is perhaps helpful to think of it as an analogous problem to the frame problem, well-known in philosophy. Although the frame problem is concerned with first-order logic, whereas the curse of dimensionality describes the problems of data analysis, there appears to be a convergence on signaling the difficulty in constraining relevant knowledge to effectively navigate an exploding space of possibilities. It is a problem of how to decide which features are relevant without the prior analysis of them all (which is obviously impossible because of their gigantic volume).

demonstrated how the neural networks perform this kind of dimensionality reduction (sometimes perhaps more technically and certainly more dramatically, the dimensionality collapse). They achieved that by mapping the statistical regularities of the environment (dataset) onto the initial weights of the intermediate layers (through pretraining). Although the technical side of that should not concern us that much, it is important to understand that we can now plausibly demonstrate how the usage of structure in the environment (regularities in the dataset) effectively constrains quasi-perceptual performance (although admittedly in isolated tasks such as image classification, so far).

The generalization of the machine learning framework for perception is still a huge 'if,' but some crucial pieces of evidence are already in place (and rapidly accumulating). That is why it is more useful to talk not of a wholesale perceptual optimization strategy, but of a combination thereof. These strategies are relative to both the stage of the perceptual processing (as in the case of whitening and sparse coding) and particular tasks. It is important to remember that conditions (constraints) on the optimality of the process are relative to how we delineate the system and assess its goals. This might seem trivial at first but should make us wary of speaking of global optimality in perceptual processing all along the (e.g., visual) stream and also incline us to analyze the compression-like powers of the particular parts of perceptual systems, respective to their functional parts, contrary to the ambitions of the likes of Wolff (1982, 1993).

To summarize, we can invoke the three principles for efficient compression in perceptual systems as recently proposed by the aforementioned Bates and Jacobs (2020:3-5). Their formulation seems to recapitulate nicely the properties we discussed so far in favor of understanding perceptual processing through compression lenses. These three principles are, in their terminology:

Limited capacity principle: which states that due to cognitive limitations of human subjects as perceivers, they are doomed to apply the generalize-then-generate kind of procedure to cope with the highly complex yet structured environment

Prior knowledge principle: which states that to compress efficiently, knowledge about the statistical properties of the signal is necessary. Their formulation in terms of actual knowledge (such a view is outspokenly defended also by Palmer [1999]) is, of course, problematic (and as I argue in a later section, while discussing the representational format of information, misleading and unnecessary). But regardless of whether we treat the statistical model as internally represented or not, or rather whether it is 'knowledge-within-the-system' or rather 'knowledge-to-the-system', we can define this condition as taking advantage of regularities in the signal while explaining the systematicity of perception.

Task-dependency principle: which states that any compression-like procedure is effective regarding a certain behavioral goal of the agent.

Now I would like to turn to some more empirical evidence of such specialized compression-like strategies being employed by the visual system.

2. Empirical evidence

We have so far briefly discussed the simple case of static image processing and mentioned that the visual landscape is significantly more complex. The brain copes by utilizing the deep regularity ingrained in the complex yet systematic environment, as evidenced by the Natural Scene Statistics paradigm. We know that pixel intensities in the pictures are not random but highly autocorrelated (Ben-Av & Sagi, 1995); the same goes for edge orientation (Geisler, 2001) and even perhaps

categorical features (Oliva & Torralba, 2006). But the picture (literally and metaphorically) gets blurrier the faster the image moves.

This intuition is widely utilized in the video-processing community. A moving picture, contrary to a static signal, presents an opportunity for two complementary compression strategies—intra-frame compression and inter-frame compression. Intra-frame compression equals roughly treating the particular frame as a static image and processing it according to static-image compression algorithms (e.g., PNG). But the inter-frame coding provides an opportunity for a more greedy approach. It makes intuitive sense that representing every next frame literally, in a sort of bitmap seriatim processing, is hugely inefficient. After all, we can group the block of frames, assuming they would not differ that much (similarly to the assumption of autocorrelation of pixels in static images—we take advantage of the fact that the natural world, with disregard to all appearance of turbulence, is really quite stable). If that is the case, we needn't represent every new frame in isolation (and in a sense, from scratch), but can only register the difference between the given one and its predecessors. This is why livestream videos containing snowfall or confetti dramatically lose resolution, whereas the static frames with them are perfectly sharp. The video processors do not work like that; they try to compute the difference (e.g., to perform a simple transformation of form, move this batch of pixels by this many pixels) or sometimes use simple memory storage (e.g., MPEG-2 format does). Of course, I am simplifying here greatly—the video coding is an extremely complex craft, and the abundance of techniques is a direct corollary of their variety. Nonetheless, we can be content with the simple observation that the capture of moving pictures creates new dimensions for structure, regularity, and predictability, which is enabled by movement. Hence, we can reasonably ask, does our perceptual system take advantage of that and compress accordingly, too?

This problem was investigated by Nortmann et al. (2015), who analyzed the reaction of the feline primary visual cortex to such rapid changes of stimuli through voltage-sensitive dye imaging. Researchers presented the subjects with sequences of images (natural scenes) and their superpositions with edges of particular orientations removed. The results were surprising. According to them, V1 actually employs two different coding strategies, depending on the time elapsed between the differing versions of the image. On a shorter timescale (30 milliseconds), the neurons represented complete image information, dubbed the 'ongoing representation.' But as the elapsed time between the sequences was extended to more than 100 milliseconds, the neurons represented only the image differences (what was new or missing, depending on the order of presentation). Such evidence for dual computation purely on such a very early level is surprising. The visual cortex suppresses redundant information. By doing that, it saves resources by frequently forwarding image differences, pretty much as the methods used for video data compression we discussed earlier.

This suppression of redundancy does more than save bandwidth; it constitutes the system's intrinsic motivation to model the world. Schmidhuber (2009) radicalizes this view, arguing that the drive for compression is not just a constraint imposed by the hardware but the fundamental objective function of the cognitive system—the engine of curiosity itself. In his formal theory of creativity, the brain is modeled as an active data compressor that seeks out data it can learn to compress. Crucially, "learning" is equated not with the static state of being compressed, but with ‘compression progress’—the first derivative of the description length.

The brain gets a "reward" (an intrinsic reinforcement signal) whenever it finds a shorter program to encode its history of observations. This dynamic explains the "Goldilocks effect" in attention and curiosity. The system ignores pure white noise because it is fundamentally incompressible (no pattern can be found, so no

progress can be made). Conversely, it ignores static, predictable walls because they are already maximally compressed (no *new* progress can be made). Attention is naturally drawn to the "frontier of learnability"—novel but structured patterns where the system can rapidly reduce the description length by updating its internal models.

Consequently, perceptual phenomena such as "filling-in" (e.g., in the blind spot, illusory contours, or the amodal completion of occluded objects) are not passive illusions but active predictions generated by this compression engine. Because the system has internalized the regularities of the visual field to maximize compression, it encodes the scene efficiently as "continuous surface + deviation." When the sensory input is silent (as in the blind spot), the generative model simply "unzips" the expected regularity. The brain does not need to "invent" the missing data; the missing data is implicit in the compressed code it has already selected. The ability to hallucinate the invisible is the direct computational yield of the drive to find the optimal code for the visible.

It is important to notice here that both in the visual compression case and in Nortmann's study, it is very tempting to say that what explains the process is some kind of prediction. Indeed, in both cases that is exactly the type of language we encounter in their analysis. This is routine in discussions of compression schema in video processing. For example, Mavlankar et al. (2006) when advocating for the particular coding scheme of their design, claim that its 'intraframe prediction provides an alternative and increases coding. The skeptic would probably be quick (and right) to object that this kind of talk is simply technical, disciplinary jargon of little philosophical interest, as everybody would agree that neither cameras nor the compression schema predict in any philosophically robust sense of the term, even if the user, for once, downloaded all the right codecs.

Let's move to the physiological case, then. The authors themselves admit predictive interpretation as the most plausible one. To quote the passage: „responses to stimulus differences fit well conceptually with predictive coding principles, proposing that deviations from cortically generated predictions are propagated up the visual hierarchy as error signals." But that seems hasty at best. The same effect can be explained by adaptive behavior. The bias inherent in the presented signal is counterbalanced by the visual system by introducing its own 'counter-bias' in responsiveness and selectivity so that the average firing rate of a given neuronal population ('population signal') was kept constant. Such a principle has long been advocated by psychophysicists to maintain equality in the time-averaged responses across the population and to enforce independence in selectivity across the population. The evidence for this homeostatic hypothesis has indeed been found by a team led by Benucci (Benucci et al.,2013).

The structure can be meaningfully described as performing information compression as it actively suppresses redundant informational intake to be sent further downstream. It is an optimizing, efficient neural-informational procedure. But it's no more predictive than the optimizing, efficient information procedure of video compression.

Compression seems to be a case of a biased process defined over informational states. But how can informational states fare better in explaining perceptual processes? In the next section I briefly consider some objections to information-based accounts and suggest how these worries might be alleviated by existing theories of naturalization of informational content. I am mostly concerned with the problem of bridging information-based principles with accounting for representational powers in perception. I argue that holding firm on the informational analysis opens the gate for meaningful explanatory mapping from artificial to natural information processors that keeps their representational

capacities but does not overintellectualize the problem like more constructivist accounts, such as predictive or inferential ones, do.

3. Informational states and representational capacities

The common charge against the information-based explanations accuses them of the inability to account for content. The informational states are supposed to be not about things as such, unable to contribute to the meaning. The orthodox interpretation of information theory seems in line with this assertion. Shannon himself was openly skeptical about the extensions of his theory to anything more than a mathematical theory of effective communication and formal description of signaling channels. He considered the ascription of meaning to information theory to be a fundamental misuse of its core function and attacked it in an article forthrightly titled 'The Bandwagon' (Shannon, 1956).

The rudimentary problems seem to be how to account for the shift from simple signaling systems to contentful representative powers of the given systems. Precisely this riddle seems to motivate the constructive account of cognition and perception itself. Predictive processing, inferentialisms of all sorts, and enactivists share the belief that the scaffolding has to come from above all the way on the ladder. In perception, this problem is chiefly motivated by what is called the 'inverse problem.'

The old venerable 'inverse problem' in philosophy of perception is arguably as classic as it gets within the discipline. The problem, which basically pertains to the troubles with explaining perceptual experience in the light of underdetermined stimuli, seems to come back in ever newer iterations from at least the Kantian times (Hatfield, 2002). The riddle is, again famously, whether perception can be explained purely by the states of the world-out-there or rather some kind of

agential interpretation is necessary for any kind of perceptual process to succeed. Obviously, the most important milestone was set by Helmholtz (1867) and his hugely influential doctrine of 'unconscious inferences' in perception that was supposed to solve the problem for good. Unsurprisingly, the establishment of the Helmholtzian tradition and its long lineage (Gregory, 1966, Marr, 1982, Rock, 1983) was only the onset of the much-prolonged trench warfare between proponents of inferentialism and its adversaries—the successors of the empiricist tradition that claimed that perception is a passive process of receiving particular kinds of information about the environment, such as Gibsonians. This venerable old debate has found new fuel and exciting reconceptualization with the advent of yet another round of active perception theory, such as Bayesian theory and predictive processing framework. Leaving aside the differences between the two, it suffices to say for now that both claim perception is necessarily involved with generating rivaling hypotheses about the incoming perceptual input, against which the proximal stimuli are then judged. These hypotheses (or assumptions) are encoded as priors and hyperpriors (the latter differ from the former by their wider scope or being of a more general nature; the information that the light is generally expected to come from above is hence an example of a hyperprior) and then used to calculate the posterior probability of the percept (given the particular hypothesis) (Orlandi, 2015).

A contrary view, representing the non-constructivist camp, rejects the full-blown inferentialist view as over-intellectualizing perception. The rival view holds that there is a simpler, yet more empirically grounded framework to explain perceptual processes that does not resort to much else than the properties of the perceptual signal itself to solve the inverse problem. This family of views is based on the Natural Scene Statistics (NSS; Oliva & Torralba, 2001; Torralba & Oliva, 2006; program in perceptual psychology and artificial intelligence that holds that it is the thorough analysis of the properties of the signal itself that can explain away a lot of

the inferentialist constructs. It was recently most eloquently and powerfully put forward by Orlandi (2014; 2015) and dubbed 'The Embedded View' (EV). The EV basically holds that perceptual processes are better understood as biased processes defined over non-representational states. Orlandi however, does not discard the Bayesian outlook but offers its major reinterpretation. They claim that Bayesian inferences are not legitimate but merely metaphorical inferences. If the inferentialist concedes that, however, his position loses any purchase as a constructivist theory, a real alternative to the ecological perception.

What is crucial to understand in this debate is that both sides agree on the fact that perception is a non-random process that heavily utilizes regularities, although they differ on the origins of these regularities. Both also seem to agree (as virtually all cognitive theorists do) that perceptual systems are in the information processing business. When one takes both of these points into consideration, it becomes quite startling that a crucial idea about the usage of regularities in the informational input has been mostly neglected. That idea is compression. One of the main points of this chapter is to argue that it is favorable to use this concept to better understand the biased nature of perception without resorting to vague and unproductive overintellectualizations. Hence, I agree with Orlandi with their general statement on the nature of early visual processes and suggest a complementary way to argue for a non-constructive account.

Let's go back to the problem of the emergence of content that purportedly disqualifies informational explanations for representational powers. The obstacle on the way is called 'the hard problem of content,' which posits that the criterion for being a contentful state is to possess satisfaction conditions. This argument has been recently refreshed by Myin & Hutto (2013; 2017) in their widely discussed argument of 'bootstrap hell' against the contentfulness of basic cognitive systems. I do not want to dwell further on their proposal; suffice to say for now that,

according to radical enactivists, informational states are supposed to merely lawfully co-vary with the environmental features without the property of being about them. Think of the number of non-decayed atoms of C14 carbon within an archaeological object (say, a jug) that lawfully co-varies with how old the sample is, but the ratio of decayed and non-decayed atoms itself is not about the age of the jug. There is information within the sample (in the Shannonian sense), but there is no bridging that with the contentful, semantic account of information. The jug surely does not represent its own age. The Fodorian case of rings on a tree is another classic example of a structure that lawfully co-varies with some property (age) but is not 'about' in a within-the-system sense (Fodor, 1999).

I think this kind of objection is misguided, and there are already a host of promising philosophical theories that suggest the way to circumvent the hard problem of content or render it invalid. Skyrmsian theory of evolutionary game-theoretic signaling systems, building on Lewis' (1969) pioneering work on signaling games, is one. Shea, Godfrey-Smith, and Cao (2017) also demonstrated how simple signaling systems can achieve states that satisfy the required condition of being contentful. A slightly different, more conceptual approach of demonstrating that Shannonian information itself is indeed semantically rich has been recently put on the table by Isaac's work on latent Shannonian semantics (2019) and Martinez's (2019a) idea that Shannon's information is equally a theory of communication and representation.

The problem was sometimes stated in terms of the possibility of misrepresentation. Informational states are not contentful because lawful covariance does not leave the possibility for the system to 'be wrong' about what it informs about. These kinds of worries were addressed by Dretske's protopredicative indicator theory (1982), Millikan's teleosemantics (1984), or the idea of asymmetric dependency criticized by Fodor (1990) by trying to offer a more minimal account of representation. However, the skeptics of minimal representation in simple cognitive/perceptual systems

remain unshaken, pointing at other criteria that ought to be satisfied before we should be inclined to admit the representational status of a particular information processor (note that label is virtually uncontroversial across the board).

One of those important criteria is decouplability. The ability to act on absent stimuli (informational intake) is yet another. A somewhat similar desideratum to the former is sometimes formulated as 'distance' from representation to what it co-varies with. Let's start with the inverse order.

“Distance is conceptualized as the capacity of representations to occur (in principle) in the absence of their typical causes. This is supposed to underscore the nature of informational states as mere causal mediators or 'relays.' States that (even asymmetrically) co-vary with something and/or have the function of doing so are too close to what they causally and statistically correlate with” (Orlandi, 2014:121).

Absence is another criterion. A handful of examples are provided by Cantwell-Smith (1996), who considers the example of coyote and rabbit. Rabbit carefully keeps an eye on the coyote (as he is one clever beast). As far as the coyote is within the line of the rabbit's sight, there is a connection between the predator and prey's perceptual system. However, when the coyote turns a corner and hides behind a rock, the previous link is broken, yet the rabbit still guides his behavior according to the position of the coyote. In a sense, now the rabbit deals with something that is perceptually absent to it. A corollary of the ability to represent absence is the ability to 'fill in' the holes in perceptual outlook. 'Completion is an important mark of the emergence of representational states. It indicates that perception allows us to track what is in the world even when what is in the world is, in some sense, not present.' (Orlandi, 2014:126)

Gestalt-like laws or constructive (inferential and/or predictive) accounts are usually mobilized to explain this important feature of successful perception. But as we have

seen, compression-like systems that thrive on statistical regularity in pattern recognition are custom-made for dealing with such biased processes just as they are—just biased.

All three intuitions—absence, distance, and completion—can be generalized to the idea that representations have to be in some way disattached from the online perceptual input. Numerous authors (Clark, 1997, Chemero, 2009, Gallagher, 2017) have indicated that this notion of *decouplability* (that we covered extensively in ch. 2; section 5.2) plays an important role in explaining the representational character of cognition.

This idea of detachment from 'online' stimulation seems to mirror an old philosophical intuition that seems to get a bit of a bad rap (or at least a label of antiquated, folk concept). Namely, the idea of a cognition/perception boundary. It is precisely that independence (in a loose sense of the word; cognition is in almost all accounts *somehow* inextricably connected to proximal stimulation) from the actual sensory activation. There are, of course, complications to that account, such as endogenous hallucinations[168], demonstrative thoughts, and perceptions of absence[169]. Beck (2015) defuses them nicely and offers a plausible definition of perceptual processes to differentiate them from cognitive ones[70].

---

[68] Think of acid-wrenched Johnny Depp getting assaulted in a bar by the reptilian envoys of the New World Order in *Las Vegas Parano.* Not to be confused with the thirsty, exhausted wanderer in the middle of a Saharan caravan who experiences a mirage of an oasis—an exogenous hallucination.

[69] An important clarificatory note is in order here. Perceptions of absence are not to be confused with absence. The latter are types of cases as discussed in the rabbit and coyote case. For example, when the absence of something (coyote) is a part of the bigger cognitive picture (the entire episode and the perception of the rest of the environment, such as the rock). Perceptions of absence are situations when an agent realizes there is no stimulation (of a given modality) whatsoever—such as when the sudden silence strikes.

[70] His criterion for being perceptual is formulated as a 'full functional stimulus-dependence': „Ψ is perceptual if, necessarily, all occurrences of all elements of Ψ have the function of being causally sustained by present proximal stimulation; otherwise Ψ is cognitive.

According to different flavors of multi-layer predictive theories, the existence of 'joint in nature' (Block, 2014) between perception and cognition has been rendered obsolete by the new framework. There are supposed to be, in fact, many more interacting layers acting on the roughly similar organizational principles (say, hypotheses tokening and testing). But as we see, the notion of decouplability seems to be just another way of sneaking the old distinction to the debate through the back door yet again.

The problem of conceptualizing the interface between perception is sometimes (most famously by Fodor) conceptualized as the 'information encapsulation problem.' For Fodor, the stake was to uphold the modular view of the mind. He (along Pylyshyn; 1999) devised the idea of encapsulation to fend off the New Look psychology (spearheaded by Bruner and Postman; Lupyan, 2016). But the debate eventually floundered with the inability of phenomenological (e.g., argument from stability of illusions) and behavioral evidence to account for definite encroachment of the modular borders (from higher to lower levels). A little semi-historical note is perhaps in order here: Fodor was extremely skeptical of the type of imaging evidence. Most famously, he lambasted the usage of 'coloured maps of brains' as tools for inferring the informational content of brains in his 1999 article for *London Review of Books* (Fodor, 1999). Most recently, Coltheart (2006) expressed similar worries that imaging evidence is next-to-useless for mapping the psychological phenomena into physiological ones. I argue we now possess better tools to take the 'informational' part of the encapsulation problem more seriously and try to present novel techniques that display promising results for such an enterprise—artificial networks as models and sophisticated imaging techniques.

4. Structural and representational similarities. Infomorphisms?

But newer techniques such as Representational Similarity Analysis (Kriegeskorte et al.,2018; Roskies, in press: for philosophical analysis) are available. The highly influential paper on the method (Kriegeskorte et al.,2008) offers a tour de force of it. The aim of the study was to present the same set of visual images (the set contained many categories of objects) to humans and macaques and register the activation of the inferior temporal cortex (through fMRI in humans and electrophysiological recordings in monkeys). Then a clustering algorithm was used to build a matrix representing the response of recorded voxels (humans) or neurons (monkeys) in response to particular categories. The analysis of similarities of the matrices discovered a similar mapping that entailed some semantic-like relations between objects (living and non-living kept separate, followed by according subcategories—e.g., living creatures, then body parts). In each case, a clustering algorithm was applied to the response of many simultaneously recorded neurons/voxels to visualize a low-dimensional embedding of neural responses to specific images. Across species there was a highly similar clustering of visual images, with a distinction between living and non-living objects at the top level, followed by subcategories within those larger categories (such as faces versus body parts, or natural scenes versus objects). That suggests RSA allows for cross-format structural comparisons and also gives hope that we can infer representational contents from proxy vehicles (recordings). Kriegeskorte does not hesitate to explicitly call the representational similarity project a 'shift from activation to information' (Kriegeskorte et al.,2008). Interestingly, Bates and Jacobs report (2020) that the use of their compression-based model (see: the rate-distortion theory we already extensively covered on the basis of Martinez (2019a), which operationalizes compression as an organizing principle for training, representational space similar to that reported in Kriegeskorte's studies emerges.

But what does it mean for two neural representations to be "structurally similar"? We can ground this intuition mathematically using algorithmic information theory.

Standard measures of similarity in cognitive science and philosophy often rely on geometric distance (e.g., Euclidean distance) within a predefined feature space. This approach is fraught with difficulty: it requires the researcher to arbitrarily define the dimensions of that space (the "Frame Problem"), and it breaks down when comparing systems with different architectures (e.g., a biological brain vs. a silicon chip) that do not share the same physical feature axes.

To tackle this, we can turn to Algorithmic Information Theory we already discussed in Chapter 1. On its basis Cohen & Vitanyi (2015) proposes a fundamental, parameter-free metric: the Normalized Compression Distance (NCD). The intuition behind NCD is that two objects are similar if they *compress well together*. Formally, if the description length of the concatenated pair (FileA+FileB) is significantly shorter than the sum of their individual description lengths ($C(A,B) \ll C(A)+C(B)$), it proves that A and B share mutual algorithmic information. The compressor can use the patterns in A to predict the patterns in B, reducing the total file size.

This metric is "universal" because it does not rely on specific features; it detects shared structure regardless of whether that structure is syntactic, semantic, or causal. This provides a rigorous foundation for the structural realism observed in representational similarity analysis (RSA). When researchers like Kriegeskorte find that human and macaque visual cortices share a "representational geometry," they are effectively showing that their internal states share the same algorithmic complexity profile. Even if the neurons (the hardware) are different, and the firing rates (the code) are different, the *compressibility relations* between the states are preserved. Thus, "structural similarity" is not a vague visual metaphor; it is a

concrete measure of shared algorithmic content defined by the universal properties of the vehicle's format.

I argue that to do that, we need to be able to reliably track information flow (and its direction) but also account for the right format, i.e., identify according representational structures. For example, inferentialism holds only if the format is proposition-like. But that is difficult to empirically validate (which perhaps explains the persistence of these theories). Thus, it seems to me that Orlandi's explicit ambition to provide a conceptual framework to disentangle constructive from non-constructive processes in vision is a well-motivated one. The difference hinges on the nature of regularities used by the perceptual system. Their idea is to introduce the differentiation between sensory representations and principle representations. The goal of that is to differentiate whether the system has the regularity encoded within itself (has the relevant *r-representation*; i.e., representation of the rule) or merely behaves in accord with that rule in a trivial sense (as Mondays behave according to the rule that they come after Sundays) (Ramsey, 2007). This is a critical point for the success of a genuine explanation of the informational and representational structure of the system. We have to clearly delineate what is information-within-a-system from information-to-the-system.

The dichotomy comes in handy, for example, when analyzing imaging studies. When we hunt for voxels and find out they regularly co-vary with a particular task of stimuli, that information that the voxel is a vehicle for may not necessarily be the vehicle (true information carrier) for the brain (Egan, 2014). On the contrary, it almost surely is not the case—voxels are artifacts and do not map nicely onto structural boundaries between brain structures. They are also somewhat coarse-grained, as each voxel is a cube of 3 millimeters containing millions of neurons. Certainly, voxels are reliable indicators of information processing within the brain, so their activation does seem to co-vary in some (much)

better-than-chance way with the brain activation. But it is unwise to assume that there is a strict one-to-one mapping or any kind of identity relation between the two.

This is precisely the guardedness that keeps us in line with the parsimony and accuracy of our theory. If we can explain Mondays, digital cameras, and perhaps V1 cortex ensembles – all of what we know about them – without the resort of high-level predictive powers, then we should do without predictive language, same goes for inferential explanations (e.g., Kiefer, 2017). We theoretically could treat them 'as if' they predicted or inferred (as it is sometimes valuable—for example, for didactic or heuristic purposes), but for explanatory purposes, why should we? That is not only confusing—implying deep connections between phenomena that are not necessarily related—but also explanatorily barren. Overextending concepts trivializes them (if almost everything performs a certain function F, it no longer tells us anything that X Fs; as we discuss information processing, we can say that as the probability of X F-ing approaches 1, the informational value [surprise] nears zero). Point in case—the notion of inference was overextended precisely for the purpose of explaining the purported constructive aspect of perception (Lupyan, 2016) or cognition more sweepingly (Friston et al.,2011). Despite the valid criticism of Orlandi (2016) that Bayesian inferences are not genuine inferences (operating on representational structures in accordance with language-like rules), the proponents of predictive accounts double down on the idea of literal interpretation of perceptual inference (Kiefer, 2017).

But it probably affords to object now that if inferences and predictions are deemed to be ascribed too liberally, why not be equally suspicious of representations? If we are advocating for an informational account, should we not try to discard representations, too? This, however, assumes that representations are just theoretical posits (as they have been a central one in psychology since its inception).

Interestingly, we might be past the point where this is the case, as convincingly argued for by Thomson and Piccinini (2018). Neuroscientists do, in fact, routinely and comfortably not only observe but also manipulate them, rendering them perfectly real scientific objects (Hacking, 1983)[71]. The neural representations come in different subclasses. Piccinini and Thomson differentiate sensory, motor, and 'uncoupled' (equaling decoupled ones in the sense discussed earlier) representations. Such real neuronal representations are to be found even in early vision, where the standard example of sensory representation of absence is the blind spot (Matsumoto & Komatsu, 2005). Sensory representations seem to be in line with the desiderata for s-representations *sensu* Orlandi.

The worry about the overly liberal policing of what counts as representation is based on the prima facie plausible assumption that 'while information is ubiquitous, representation is not' (Orlandi, 2014: 105). That should make us wary of equating informational processes with representational ones. Granted, not all information processing systems can represent. Moreover, almost none that trade in information do reach the status of truly (cognitively) representational systems. But ever more plausibly, some do. And for sure—all truly representational systems do trade in information, too. The question is, do they represent the virtue of attaining a particular information processing structure?

Quite famously by now, artificial systems such as deep neural nets rapidly achieve impressive success in a wide range of perceptual tasks, such as image categorization, image generation, classification, etc. The rapid pace of proliferation of new architectures and techniques might imply the belief that, in fact, many diverse systems can realize many perceptual functions. But that is misleading. Cao

[71] To be precise, according to the authors, this is by no means a novel phenomenon in neuroscience (they claim it to be at least a century and a half old, reaching back to Hughlings Jackson's experiments on motor representation). They simply claim that philosophy has been extremely sluggish to catch up.

and Yamins (forthcoming) observe that, in fact, only simple tasks have many solutions. More complex tasks are by definition more constrained (have to satisfy many conditions at once and trade off between them efficiently). The more constrained the task, the fewer possible solutions, hence the more convergent particular architectures (evolved or trained) to solve it become. They call it *a contravariance principle* (Cao & Yamins, 2024): as the complexity (constraint) of a task increases, the number of viable architectural solutions decreases (converges). Thus, complex problems force disparate systems to adopt similar structural solutions. I previously analyzed the assumption that biological systems are constrained to perform efficient coding (Barlow, 1978; Chirimuuta, 2014). That is precisely because of the highly constrained nature of their performance. If we accept it as a premise, we can see why we should, in fact, expect the artificial and biological nets to converge on similar solutions (provided we constrain them similarly).

That should raise our conviction in the fact that artificial nets do in fact provide good models for biological processes and that information-based explanations of perceptual processes would result in too many systems counting as representational. This perhaps also illuminates why the Moravec paradox happens to be the case. The intuitively simple perceptual task lies low only when we consider the ladder of abstraction. The conclusion that corresponds to the level of complexity is illusionary. The contravariance principle explains why these deceptively simple tasks are in practice extremely hard—precisely because they are in fact much more constrained.

There is a markedly strong support for the important informational similarity between biological and artificial nets (convergence) with the findings of the ability of neural nets trained on visual classification tasks to be able also to predict neural (biological) activity with relatively high accuracy (Yamins & DiCarlo, 2016).

Moreover, Bashivan et al. (2019) developed an approach called „deep image synthesis" that is used to use machine learning systems to *generate* patterns of visual stimulation that maximally activate individual neurons in a monkey's cortex (V4). The performance of the model was so high that it was able to generate stimuli that resulted in rates of neural firing in these neurons that were far greater than those evoked by any available natural image stimulus. (Poldrack, 2021). A similar idea of identifying the interpretable building blocks that maximally activate a particular neuron to explain its function is also making a significant impact in 'explainable machine learning' (Olah et al.,2018).

What's equally important in our ability to localize and manipulate (e.g., through increasingly potent optogenetic or ablational techniques) neural representations is the idea that we seem to start being able to compare particular representational structures between different agents (Firestone, 2020). Two especially impressive examples come from Schalk et al.'s (2017) and Marshel et al.'s (2019) studies that seem to be able to optogenetically 'incept' a certain (carefully designed) representation within the brain. The former study was able to invoke face-like percepts in human subjects overlaid on inanimate objects. The latter study achieved successful substitution of visual stimuli that guided mouse behavior without the loss of ability to perform a high-order function.

We have analyzed evidence from conceptual analysis, similarity analysis, and neuronal manipulation from both human, animal, and artificial systems. We have observed that this data discloses striking functional and structural similarities, enabling not only comparison but also mutual prediction. It points to the structural similarities of the different kinds of agents. We seem to be zeroing in on resolving the dilemma offered by Ramsey in his discussion on the criteria of effective representation: whether 'are there mindless systems in which an internal element is performing a role that is most naturally (or intuitively, or justifiably, or

beneficially) viewed as representational in nature?' (Ramsey, 2007:196). Recent advances in the successful mutual comparisons and predictions from diverse (artificial and biological) systems lend support to the idea that there are deep structural similarities that explain the converging representational patterns as optimal coding procedures that effectively maximize the 'output-to-scaffolding' ratio, very much compression-like. Perhaps that is plausible evidence for the existence of structural similarities not only between the environment and the agent but also between the agents themselves.

Structural similarity is an important feature for many theorists for explaining the representational capacities of the system (Cummins, 1996). They successfully represent because there exists a systematic morphism (iso- or homomorphism) between important parts of the environmental makeup that are mirrored within the agential makeup that is in the business of faithfully representing its surroundings. This relation can be perhaps fruitfully described in the form of infomorphisms, as suggested by Barwise and Seligman in their 'Information Flow: Logic of Distributed Systems' (1997). Perhaps the notion of 'structural similarity' (Gładziejewski, 2015) between the environment and the representational system and the idea of S-representation (Gładziejewski & Miłkowski, 2017) can be mobilized for this purpose. This is a promising line of research, but so far—the one for the future.

5. Conclusions

This chapter was concerned with how artificial and biological systems share not only the feature of being information processors but also the challenge of performing this processing in a highly selective, regimented way. This kind of outlook on the environment as structured and, as such, predictable (and the informational reservoir it constitutes), confirmed by the findings of the Natural

Scene Statistics paradigm, invites thinking of perception as compression-like. I argued that such 'clever' strategies are not only beneficial for optimizing but also necessary from both ecological (biological constraints) and computational (curse of dimensionality) standpoints. Any successful information processing simply does not afford the luxury of total tracking. I also briefly reviewed the potential objections to informational accounts as lacking content and worries about the unstable status of the notion of information itself (especially in neural contexts). It was conceded that although the objections of skeptics are well-motivated, they can be alleviated by existing theories of naturalization of informational content. It was argued that the postulated necessity of taking advantage of regularities in conjunction with content has to be interpreted representationally in the light of recent riveting breakthroughs in observability and manipulability of neural representations and their computational modeling in artificial neural nets. Finally, I tried to demonstrate that the ability of mutual modeling and prediction between biological and artificial neural structures points to the possibility of establishing a structural similarity based on the notion of infomorphism.

Chapter 7: Meaning from Matching: How Representational Mechanisms Exploit Structural Similarity[72]

1. Introduction: From semantic information to representational geometries

In Part 1, we established the core commitments of the Neo-IPP. Among these is the strong commitment to decodability: if the cognitive system is truly trafficking in contentful representations, and not merely dynamical couplings, then this content must be recoverable from the physical substrate. This brings us to the empirical frontlines of cognitive neuroscience. We must subject our framework to a rigorous stress test: can it explain the success of modern decoding practices, or does it collapse when faced with the raw data of the brain?

---

[72] This chapter is based on the article under the same title (published in *Theoria* in 2025, co-authored with Marcin Miłkowski). To adhere to formal requirements regarding the inclusion of co-authored work in doctoral dissertations, sections primarily drafted by the co-author—specifically the section exploring the connection of the CNT to mechanistic explanations - have been omitted. The remaining material, representing my distinct contribution, has been repurposed and augmented with previously unpublished arguments to serve as a targeted empirical stress test for the Neo-IPP framework defended in this thesis. Thus, this version of this chapter is more limited than the published version (Mamak & Miłkowski, 2025). An inquisitive reader is more than welcome to consult the 'fuller' version in the case of insatiability (if that's possible while engaged in absorbing a 300+ page dissertation).

While Chapter 6 demonstrated that the format of representation is driven by the imperative of compression, we must now ask how the content of these compressed structures are fixed and shared across different substrates. If compression explains the "how," structural correspondence explains the "what."

To execute this test, we turn to Representational Similarity Analysis (RSA). Since its introduction, RSA has revolutionized neural decoding by shifting focus from the localized activation of voxels to the geometry of representational spaces. Crucially, RSA leverages the principle of medium independence: it allows researchers to compare data from disparate substrates—human fMRI, macaque single-cell recordings, and artificial neural networks—by abstracting away from the physical implementation to the relational structure of the information.

However, the success of RSA raises a profound theoretical problem, often termed the Decoder's Dilemma: how can we distinguish between genuine f-information (functional information-for-the-system) and mere a-information (statistical artifacts or information-about-the-system observable by the scientist)? How do we know that the "geometries" we detect are the actual vehicles of thought and not just spurious correlations?

2. The Correspondence Network Framework

This chapter argues that the solution lies in the Correspondence Network Framework (CNT). We posit that RSA works precisely because it detects networks of infocorrespondences—structural mappings that hold between the environment, the neural vehicle, and the downstream consumer. It is this structural correspondence that enables the alignment of f-information and a-information. When the scientist’s

decoding model matches the system's internal structure, it is because both are tracking the same stable, structural features of the signal.

This alignment explains the phenomenon of portability. As we will demonstrate, representational geometries are not arbitrary; they exhibit specific property retention and clustering (e.g., taxonomic hierarchies) that persist across different physical media. This clustering is the fingerprint of fixed content. For a mechanism to be representational—in contrast to the broad, liberal definition of "mental mechanisms" offered by Bechtel (2008)—it must do more than process information; it must exploit these preserved structural clusters to guide action.

By analyzing RSA through this lens, we show that "representational geometries" are the correctly formatted informational structures required by the Neo-IPP. They prove that content is not a "gloss" added by the observer, but a structure-preserving mechanism used by the system to represent efficiently.

3. Representational Mechanisms Exploiting Correspondence Information

The account of correspondence-based information, summarized above, does not provide any detail on how this information may be exploited by cognitive mechanisms. In this section, we argue that correspondence-based information can be exploited by a class of representational mechanisms. These two frameworks, joined together, provide a formal and causal account of structural representation, which is usually presented as providing a crucial feature: it legitimizes surrogate reasoning between structures. From the preceding, it should be fairly clear that infocorrespondence does the same because of the structure-preserving nature of

this relationship. Representational mechanisms can thus leverage it to fulfill their key functions.

Specifically, the informational structure referred to by a mechanism exhibits infocorrespondence with a target structure. This licenses the identification of characteristics of the target based on the information vehicle. Evaluating information also relies on assessing similarity between the structures to determine accuracy. Vehicular discrepancies reveal errors in representing the target.

The mechanistic approach demands a form of realism concerning representation, implying that both the vehicles and the contents play a causally relevant role in the phenomena governed by a particular mechanism. This emphasis on the causal relevance of both vehicles and contents is one of the conditions for representational mechanisms dealing with structural representations. Further conditions follow Gładziejewski's (2015a, 2016) widely accepted contemporary philosophical analysis of structural representations, which can be summarized as follows. In the context of a system denoted as S, a vehicle denoted as V qualifies as the vehicle for a structural representation of a target entity, designated as T, if and only if the following conditions hold:

(1) Structural Similarity: V exhibits structural similarity to T.

(2) Action Guidance: The structural similarity outlined in (1) enables V to provide guidance to S's actions concerning T.

(3) Decouplability: Condition (2) remains applicable even when V is detached or decoupled from T.

(4) Error Detection: S possesses the capability to detect representational errors generated by V.

The present account endorses conditions (1), (2), and (4), while rejecting condition (3), at least in a strong interpretation. While there may be decouplable representations in representational mechanisms, the fact that a given representation remains coupled (for example, through perceptual inputs) to the target does not disqualify it from its representational roles (Miłkowski, 2017a). The condition (2) implies that representational mechanisms should guide action in a broad understanding of the term. What should be clear, however, is that the same vehicle V can guide various actions when exploited by various downstream mechanisms. Indeed, as contemporary research suggests (Mante, Sussillo, Shenoy, & Newsome, 2013), the same vehicle, depending on the computational processes it is involved in, can be processed in various ways to derive distinct downstream representations. In fact, this could be a computationally efficient way to make various kinds of use of the same input information. At the same time, vehicles attain their status as vehicles of structural representations only if they can be checked for consistency with other vehicles, which enables error detection in cognitive representations (Bielecka & Miłkowski, 2020).

As Miłkowski (2013) notes, this makes some discussions about various extrinsic content ascriptions moot: for example, notorious bug detectors may not fulfill any representational role just because amphibians in general do not seem to detect any inconsistency between the signal they receive from the tongue when they consume dummies as opposed to bugs. If they do, then the exact kinds of error detectable by their representation-consuming devices must constrain our content ascriptions. Otherwise, these content ascriptions boil down to epiphenomenal theoretical glosses.

Structural similarity is understood here in terms of the correspondence-based account of semantic information—it must conform to the conditions that infocorrespondences fulfill. This implies that there must be two classifications of

tokens into types that are interlinked with at least a partial mapping. As a matter of fact, the reason why this linkage exists in cognitive systems is that these systems respond to information vehicles systematically, establishing an information channel, which, in turn, triggers the information flow. Systematicity is precisely what enables utilizing cognitive information about members of the class beyond what is currently accessible to the cognitive system. Hence, one can reason about absent, past, or counterfactual tokens of the given type in a veridical and regimented way. This implies that representational mechanisms explain phenomena that are not confined to the spatiotemporal boundaries of a single individual.

A critical feature of our framework is that representational mechanisms typically exploit not just single correspondence relations, but networks of correspondences. Neural vehicles often participate simultaneously in multiple classification schemes, with different typing relations operating concurrently (Mante et al., 2013). For instance, the same pattern of neural activity in the visual cortex might be classified according to shape properties in one processing stream and motion properties in another, establishing multiple correspondence channels with different aspects of the environment.

This network framework elucidates three crucial features of representational systems. First, it accounts for how the same vehicles can carry different contents in different processing contexts—as observed in the multiple processing pathways in vision (Milner & Goodale, 1995; Pitcher & Ungerleider, 2021). Second, it enables cross-channel evaluation, where information from different correspondence channels can be compared to detect inconsistencies and evaluate accuracy. Third, it suggests why neural representational spaces consistently organize along semantic dimensions—the network preserves taxonomic structure across processing channels.

The Correspondence Network Framework thus goes beyond simple structural similarity accounts by explaining how multiple concurrent correspondence relations establish a rich representational architecture capable of content-sensitive processing.

Facchin (2024) has recently mounted a second critique of structural representationalism. His critique takes an opposite approach to Kohar's strategy. Instead of criticizing the promiscuity of ascribing structural representations, he embraces a pluralistic understanding and argues that representationalism, even in its pluralistic form, is too narrow to address its own desiderata, e.g., the job description challenge (Facchin, 2021).

According to his account, there are two distinct issues concerning structural representations: their status and their content. The former concerns whether a given physical vehicle achieves representational status, while the latter assumes representational status and questions whether the content of a representation is truly responsible for the involvement of a given part of the cognitive system. In other words, the status of structural representations is distinct from their causal efficacy.

While Facchin acknowledges that his aims are primarily clarificatory and taxonomical (he proposes four tentative taxa), his goal is not to pave the way for a more sophisticated representational account but to argue for discarding it altogether. For him, disentangling the commitments to status and content in structural representations reveals representationalism's inability to provide satisfactory answers. His main argument, though tacit, can be reconstructed as follows: charting a map of representationalist accounts does not clarify or bolster their case but instead exposes their lack of explanatory scope.

Our account, however, aims to explain how a framework based on structural relations—namely, similarity—can provide a complex but consistent picture of structural representations. While similarity is a crucial factor, it is not sufficient to explain representational phenomena. We demonstrate how content remains essential to explaining the causal powers of representational mechanisms, while also emphasizing the required elasticity of concurrent infomorphic mappings. This approach aligns with a pluralistic reading of structural representational posits, with which we agree with Facchin. As such, our proposal builds on type-token theories of content in semantic networks (e.g., Shea, 2007) but expands them by explaining how semantic, informational relations are established without remaining rigidly fixed. This extension helps address anti-representational worries in a novel way.

Now let us turn to how RSA conceptualizes similarity in terms of representational geometry.

4. Unlocking Meaning from Matching: The Role of RSA in Cognitive Science Practice

Representational Similarity Analysis (RSA) provides a compelling case study for understanding how the Correspondence Network Framework applies to cognitive neuroscience practice. RSA reveals patterns of correspondence across multiple classificatory schemes, detecting networks of similarity relations rather than isolated structural mappings. This methodology implicitly relies on the taxonomic organization that emerges from correspondence networks, making it an ideal testbed for our framework.

RSA is a family of multivariate pattern analysis methods employed in neuroscience to delineate 'representational geometries' of cognitive systems (Kriegeskorte &

Kievit, 2013). Intuitively, representational geometry is expected to offer a systematic map of how the perceptual state space is organized. Brain activity patterns, elicited by various stimuli, are measured (notably, an array of methods, from single-cell recordings to wide-area fMRI, can be applied), and correlation between these responses is calculated. Strictly speaking, the objects or stimuli do not participate directly in this operation, i.e., the correlation is not measured between the stimulus and the response but between two stimuli-induced responses that differ in some experimentally interesting way.

These comparative results are subsequently compiled into a systematic numerical description of the extent to which neural activity is shared across stimuli. When patterns exhibit structural similarity, it implies shared underlying representations as the source of identifying characteristics, while pattern mismatches, on the other hand, suggest representational differences. The outcome of this procedure is the description of the representational geometry. However, the precise construction of this geometry and its alignment with our understanding of semantic information raise pertinent questions.

From a technical standpoint, representational geometry serves as a mathematical description of the similarity relations between reactions to objects (scil. stimuli-induced patterns of activity) within a given dataset (Mur, Bandettini, & Kriegeskorte, 2009). The primary goal of this procedure is to operationalize the concept of similarity among these objects. These objects are conceptualized as points within a high-dimensional space, where each dimension corresponds to the range of activity measured in a specific part of the overall activation space. For instance, in the context of fMRI, this activation space encompasses all the voxels under observation, with each dimension representing the range of activity in a single voxel. These activity patterns intuitively diverge for distinct classes of objects, positioning them in different regions of this geometric space.

To quantify the degree of dissimilarity (or similarity), a similarity metric must be selected. Options include angular, Euclidean, and Cartesian metrics. The choice of metric may be driven by theoretical considerations that preserve different types of properties, although the technical details need not concern us here. It suffices to acknowledge that the procedure maintains its character as a representational space as long as the chosen metric is applied consistently.

Once the measurements are collected, the space is established, and the chosen metric is employed to compute the similarity relations, one can assert that these geometrical relationships represent mappings of semantic relations between the representations of objects. The complete set of these relationships collectively constitutes what we refer to as representational geometry. Consequently, we can view it as an operationalization of the similarity relations among objects within the analyzed dataset.

Significantly, RSA as a method is a two-step process. The initial step involves the construction of the representational geometry. To achieve this, pairwise similarity patterns are gathered. To be more precise, these patterns are derived from the subtraction of activity patterns of the same set of units (notably, in the case of fMRI, these are voxels; however, RSA is versatile and can encompass EEG electrodes, single-cell recordings, behavioral judgments, or unit activations in artificial neural networks). In a simplified example, we present the subject with stimuli such as a banana and an armadillo, record the activity across the selected units, and then compute the dissimilarity of the responses. This is calculated by subtracting the correlation score from 1 (i.e., score = 1 - correlation). These individual dissimilarity entries are incorporated into a matrix, creating a dissimilarity score. The collection of these pairwise dissimilarities between objects constitutes the first-order RSA, often referred to as a representational dissimilarity matrix (RDM). In essence, it constructs a representational geometry for the chosen objects and units.

The second step of 'actual' RSA is focused on directly comparing these first-order RDMs, aiming to obtain a measure of similarity not between pairwise activation responses but across the overall similarity of entire representational geometries. RSA, fundamentally, entails a second-order comparison or, in another sense, a meta-similarity measurement. The essence of this procedure lies in unveiling structural similarities within the internal architecture of datasets, and more fundamentally, within the first-order geometries.

While RSA is designed to capture and compare internal structural relations, this should not imply that all internal structures of the targets are fully retained in the resulting RDMs. On the contrary, we must recall that the entries in these matrices are simply numerical values, or distances between vectors in mathematical terminology (Roskies, 2021). They represent scalar quantities and, as such, abstract away from details such as part-whole relations within the considered objects or the specific presentation format. Spatial relations are also not preserved in this manner. What RSA does retain is an overall similarity score based on the chosen recording method and metric. Typically, RSA is employed to detect dataset-level internal structure, rather than focusing on the object-level structure itself. Semantic relations are defined over representations of objects and not their individual components.

RSA does not operate on the same representational substrate as the consumers within brain tissue—whether these consumers are brain regions, cortical layers, psychological modules, individual nodes, or other neural entities. The voxels typically analyzed by RSA are far more granular than the leading candidates for neural processing units. However, this discrepancy does not undermine the method's predictive success, as has been tentatively demonstrated (Thornton & Mitchell, 2018). Nevertheless, RSA, emerging from early multivariate pattern analysis (MVPA) procedures like the classical setup presented in Haxby et al. (1991),

is promoted by its designers as a means to reveal the actual information flow within cognitive systems (Kriegeskorte, Goebel, & Bandettini, 2006). Consequently, it firmly positions itself within the domain of neural decoding.

Moreover, RSA was specifically designed to address criticisms that previous MVPA methods function as spurious data classifiers when used for reverse inference. The principal objection to inferring experiential content from neural activity was that such methods abstracted away from internal relations within the signal (Rathkopf, Heinrichs, & Heinrichs, 2023). Earlier multivariate analyses failed to acknowledge that justifiable inference requires understanding the internal structure of a system's representational space. Ritchie et al. (2019) labeled the unfounded belief that neural decoding could proceed without these structural insights the “decoder's dictum,” which they demonstrated to be false. Their central question—if we cannot identify what information our classifier uses, how can we verify it operates on principles similar to the system being modeled?—reflects a specific instance of the broader problem of opacity in machine learning, where classifiers routinely take computational shortcuts (Zednik, 2021).

Kriegeskorte and his colleagues suggest that RSA is the preferred approach to sidestep the pitfalls associated with the decoder's dictum and to attain more accurate neural decoding than previously employed methods. The underlying rationale for this assertion lies in RSA's capacity to represent the internal structure within a model, grounding the psychological model used within a given system in measurable data. This, in turn, mitigates the risk of potentially superficial classification, a concern that plagued earlier curve-fitting models. What truly sets RSA apart is its unique ability to map the semantic structure by reconstructing the cognitive space, giving form to the representational geometry of the system. As contended by Kriegeskorte and his team, the RSA project serves to decode the information within a system, effectively working for the system itself, as opposed to

acquiring information about the system as external observers conducting experiments.

Upon closer examination, this approach doesn't seem to advance us any further toward the desired goal than earlier MVPA attempts. It's worth recalling that the flexibility in the types of entries accepted for second-order comparisons, along with the subsequent format-agnostic nature, casts doubt on the notion of these scalar quantities serving as genuine carriers of content. The representations within matrices of 'representational' similarities only qualify as representations in the mathematical sense of the term, as seen in the AI literature, and not as representations in the cognitive sense. Roskies (2021) aptly characterizes them as 'provisional' representations, and in our framework, this would qualify as mere semantic information. This would not be a significant issue if a more foundational, shared framework could be established in terms of information theory.

Nonetheless, the entities involved in transmitting the signal, including the vehicles and the format of the presumed coding scheme, are unlikely to directly align with the actual mechanisms operating in the brain. So, the information, in this context and scheme, remains provisional. Perhaps our present empirical approximations are sufficiently reliable. Therefore, as we progress towards more advanced brain imaging, it is conceivable that the proposed mappings could remain accurate. This would suggest that both provisional representations and provisional information may indeed be precise and mutually aligned. However, the overall empirical accuracy of RSA neural decoding depends on whether the information within and about the system eventually harmonizes.

So, is our conclusion merely critical? Are we dismissing RSA as a method for uncovering neural content? Not at all; quite the opposite. To appreciate its value, we suggest the need to:

1. Focus on its insights into semantic clustering.
2. Reinterpret what RSA reveals about neural information flow and neural information more broadly.

To address the former, we will briefly examine research on the emergence of nested classifications within RSA analysis and its taxonomic characteristics. For the latter, we propose considering information in terms of structural similarity as a correspondence relation, aligning with the correspondence-based theory of semantic information.

One of the most intriguing aspects of RSA-based studies on human and animal brains is the emergence of a categorical structure within representational geometries that exhibits reasonable invariance across subjects and even species. For instance, human and macaque inferotemporal cortices (IT) appear to follow similar coding principles (Kiani, Esteky, Mirpour, & Tanaka, 2007). The invariance being discussed is not absolute; the response matrices in the representational dissimilarity matrices (RDMs) do not precisely mirror each other's entry values. However, there is a substantial degree of overlap in response patterns between humans and monkeys.

What do we mean by "categorical structure" in this context? It implies that within representational geometries, grounded in measurable data, we identify clusters that align with semantically interpretable terms, reflecting a nested hierarchy of objects. For example, studies show a clear division when subjects view inanimate and animate objects (Jozwik, Kriegeskorte, & Mur, 2016; Kriegeskorte, Mur, & Bandettini, 2008; Mur et al., 2013). High similarity judgments are consistently observed within pairs of animate objects and within pairs of inanimate objects, but not across these categories.

This observation is also evident at a finer level of categorization within subcategories of animate and inanimate objects. For example, a sub-cluster emerges from humans to animals, followed by a distinction between quadrupeds and non-quadrupeds. Faces cluster with other facial features, while hands and limbs also group together. Importantly, this difference in the IT areas appears to be rooted in categorical distinctions rather than feature-based mechanisms. Jóźwik et al. (2016) argued against the feature-based mechanism, as their statistical analyses showed that feature-based models do not sufficiently account for the variance in the dataset, while categorical models offer a much better fit.

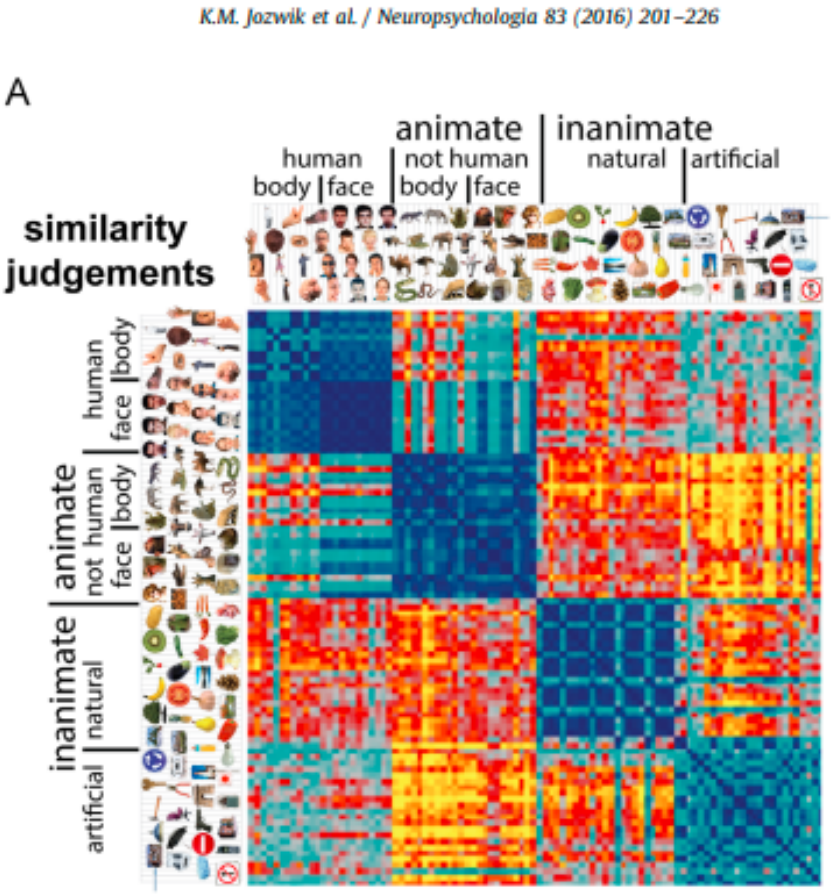

Figure 10: A diagram from Jóźwik et al., 2016 demonstrating the emergence of a nested hierarchy of semantic categories that mirror the taxonomic order of high-level features.

Why are these results of particular interest to us? In most classical interpretations (for more nuanced interpretations, see Isaac, 2019; Mann, 2020), information theory in neuroscience typically aligns with the Shannonian perspective of coding. In this view, coding is arbitrary, meaning that the content of symbols doesn't inherently correspond to the physical attributes of the entities carrying that content. While mappings rely on the statistical properties of the signal to achieve optimal coding, they do not depend on the similarity between the entities or between the intended targets and the codes. For example, in computer storage, memory addresses do not necessarily share similarity relations; they are purely symbolic.

Given the importance of semantic, similarity-based relations as organizational principles for representational geometries, it is more plausible to seek a theory to explain this property. The correspondence theory of semantic information provides such an explanation, as it takes similarity as a prior for why certain states are informative of other states. Thus, rather than conceptualizing RSA-type methods as detecting classical Shannonian information channels, we propose an alternative view: representational geometries maintain similarity-based information channels.

The RSA procedure is thus conceptualized as a series of infocorrespondence relations. First, we depict the information flow charted by the RSA analysis and the preservation of informational properties within a cognitive pipeline (see Fig. 4). This should help the reader visualize the nature of infocorrespondence relations established within the RSA pipeline.

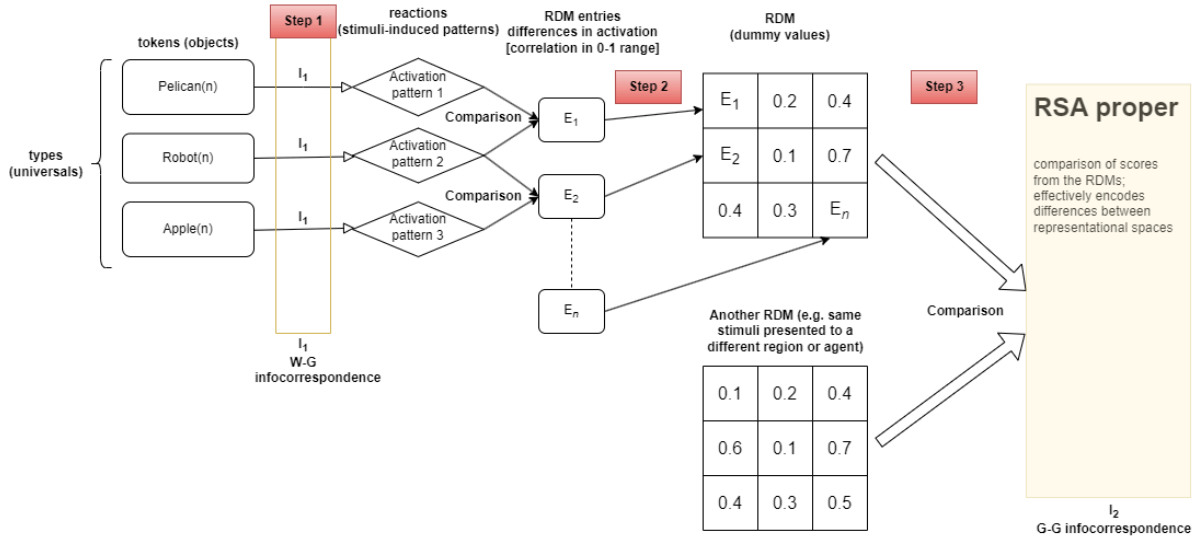

Figure 11: The RSA pipeline considered from the perspective of information flow.

In our proposal, there are three 'steps' effectively bound by two major infocorrespondence relations (I_1 and I_2). Moving from left to right, the first is the domain of objects presented as stimuli to agents. Tokens of objects are assigned to their respective classes (universals), governed by the classification operator (⊨universal). This constitutes the 'left' half of the first infocorrespondence (I_1) as they are pairs of classifications in Barwise and Seligman's terms. Tokens presented as stimuli elicit particular activation patterns, which constitute the respective tokens on the 'right' (agentive, e.g., neural) side. These tokens are likewise subsumed under counterpart types, according to the categorical operator (⊨category). Entries in the RDMs encode pairwise differences between the objects (operationalized as distances between vectors that encode these objects). It is plausible to assume that the differences are judged according to the properties that the objects do not share. Hence, at this level, these properties are the types,

whereas the activation patterns are tokens into which they are distributed. That constitutes Step 2 on the pathway.

The 'world to geometry' infocorrespondence (W-G) is crucial for disambiguating objects and placing them in proper categories. Objects lend their properties to be mirrored and utilized by agents, binding targets to representational geometries and involving actual external objects. This aligns with the view that neural information is selective responding to stimuli, with certain firing patterns reliably correlating. The binding is arbitrary, as any content could be coded by any neural state, similar to memory addresses in digital computers. However, our view diverges from this picture in later stages.

The final step—Step 3—is the actual G-G (geometry to geometry) comparison, where two full-fledged representational spaces are assessed against each other. We can then reliably infer the contents and properties of one based on the other and make 'predictions' (on the special status of the word in neural decoding, see Appendix to (Kriegeskorte & Bandettini, 2007)).

The details of the mapping may be better understood using a table (see: Table 4 below). We can see which items during the entire procedure are conceived as tokens and which as types, and what type of categorical operator binds them (governs how the tokens are classified into relevant types). A pair of infocorrespondence relations—first, between stimuli and the given representational geometry (I_1), and then, between two representational geometries that are being computed (I_2)—renders the information flow with structure-preserving (semantic) properties not only possible but also valid.

	Step 1: Stimulus exposition	Step 2: RDM – world-geometry infocorrespondence	Step 3: RSA proper – geometry-geometry infocorrespondence
Tokens	singular objects	activation differences (distances)	points in representational space
Types	universals	differentiating properties	semantic clusters or taxa
Operator	⊨(universal)	⊨(category)	⊨(σ B)

Table 4: Tokens, types, and operators inherent in infocorrespondence relations (IMs) within the RSA pipeline.

In the first column of Table 4, the classification is relatively simple. Singular objects in the world pertain to respective universal classes. Instances (tokens) of objects are assigned to their labeled sets, or in other words, universals. This is regulated via the operator we call ⊨(universal). In the middle column (linked with the first one through the first infocorrespondence relation I_1, which we denote by the pair of arrows on the top), when the objects are presented as stimuli, they are encoded by activation patterns, which in turn are translated into vectors in the state (activation) space. The distances between the vectors representing the activation patterns are compared. These distances are representations of differences between the given objects. In our terms, these differences effectively represent discriminative properties of the types of objects (types from the left column). That is why these

discriminative properties constitute types on the second step of the infocorrespondence pathway. As objects correspond to the corresponding activation patterns, universals correspond to the discriminative properties that set them apart. We call the operator working on this level the ⊨(category) operator, as its job is to assign the activation patterns into categories that they encode for the agent.

Finally, in the third step, where the 'proper' RSA comparisons are done, we encounter a geometry-to-geometry (G-G) infocorrespondence, where semantic informational relations between the representational spaces themselves are obtained. Tokens to be categorized at this level are points in the given representational space that stand in for particular targets. Types to which they are assigned are semantic clusters or taxa that contain information about the objects, but also their relationships with other objects encoded in the space (proximity stands for similarity, but certain dimensions stand for properties). We call the operator on that level ⊨(σ B). It classifies the tokens based on the structure of relations that describe/organize the representational geometry.

To wrap up, the results derived from RSA seem to imply that the coding principles within neural systems align with semantic properties—in plain terms, they convey information about what they encode. The retention of a semantic hierarchy, preserving taxonomic internal structure and clustering of mutually similar targets, serves as evidence for this architectural pattern. Current theories of information encounter challenges when attempting to explain this characteristic of 'informativeness' within a system, at least at first glance. The framework proposed here, however, does not grapple with this problem and provides an explanation for why similarity organizes the informational structure of neural systems in a semantic manner. While the mere presence of clustering could be attributed to

efficiency considerations within the traditional framework, it fails to offer an understanding of why such clustering should align with semantic categorizations.

Let's revisit the key points regarding RSA and its connection to semantic information, grounded in structural similarity:

T1: There exists an infocorrespondence relation, both among the representational geometries themselves (G-G) and between geometries and the external world (G-W) on a pairwise basis.

T2: This very relation is what facilitates the flow of semantic information.

T3: The presence of semantic information offers a more robust explanation for the retention of property relations between representational geometries, a challenge that the theory of arbitrary coding (i.e., Shannonian) finds difficult to address.

T4: Structural similarity provides the basis for researchers (external observers) to make surrogative inferences.

Importantly, exploiting correspondence information necessitates additional organizational constraints. Mere resemblance is insufficient for representation (Goodman, 1972). However, the theories that rely on resemblance seldom, if ever, ground it solely in structural relationships (Decock & Douven, 2010). However, when integrated into teleological representational mechanisms, it fosters semantically evaluable content. These mechanisms introduce further conditions to delineate suitable informational structures and targets. Correspondence, in this context, serves as the fundamental relational framework connecting them.

The theory of representational mechanisms additionally requires:

T5: Representational content is exploited downstream, which implies that representational geometries should have a causal impact on evaluation mechanisms that detect error.

This implies that the theory offers a potent heuristic: when you suspect the presence of a representation within a cognitive system, direct your attention to the evaluation mechanisms and study the subsequent processes. To put it differently, what renders RSA an incomplete account of representation is its limited capacity to systematically address this question. Nevertheless, when coupled with the theory advocated here, RSA can be employed to propose additional experimental studies.

The systemic mechanistic perspective[73] sheds light on why RSA yields insights into clusters, laying the foundation for representationally valuable inferences. Structural similarity serves as the basis for drawing conclusions about representational content and targets. Moreover, the correspondence theory of semantic information implies that as long as stimuli evoke activity patterns, these establish an information flow, which implies that there is an infocorrespondence between stimuli and activity patterns.

In summary, a systematic analysis of RSA highlights the role of correspondence-based information in cognitive systems, showing how semantic properties support mechanistic explanations and fulfill representational needs. Hence, we argue that the Correspondence Network Framework deepens our understanding of cognitive science practice and computational representation processing. Philosophically, it provides theoretical foundations for existing RSA methods, linking informational vehicles, contents, and targets—addressing a critical gap in the literature. By integrating resemblance, teleology, and mechanism, it offers a powerful framework for understanding representation.

---

[73] Developed further in the published version; see: Mamak & Miłkowski, op. cit.; esp. ch. 2-3

5. Potential objections

Now, we address potential objections to our account. Although some objections have appeared in the literature, we present them here to underscore their relevance to semantic matching.

Now, we turn our attention to addressing potential objections to our account. While some of these objections have previously emerged in the literature, they are presented here to underscore their relevance to semantic matching. There are three prominent counter-arguments that, in our view, could pose substantial challenges to the position we advocate. These objections can be categorized as follows: i) the differing formats argument; ii) the psychological models argument; and iii) the epiphenomenal correlations argument. At the end, the charge of no novelty is analyzed, too. We will tackle them one by one.

Differing Formats: The information acquired through decoding procedures, such as MVPA methods, is presented in a format that diverges from the inherent format of brain information. This discrepancy stems from the evident mismatch between the units of analysis, like voxels in fMRI research, and the actual brain tissue's intricate structure. While this misalignment may seem self-evident, its implications are far from trivial. Critics, including Ritchie et al. (2019), contend that it raises concerns about the potential confusion of different types of information, as what we obtain from these clean-cut cubes doesn't represent the formats employed within the brain.

Consequently, the information derived from these formats doesn't undergo analogous processing by any downstream mechanisms. This divergence occurs because there's no real “downstream” counterpart to voxels; they essentially cut across the natural divisions present in the brain, disregarding its authentic informational landscape of senders and receivers. In effect, the information we

acquire assumes the role of external information about the system rather than internal information within or for the system itself. There's a legitimate concern that this mismatch might result in our understanding not being expressed in the same language as the one employed by the brain (Facchin, 2023). This linguistic discord pertains to the very format of information.

It's indisputable that our methods do not employ the same informational format as the brain. Moreover, it's beyond contention that voxels and other measurement units are, to a certain extent, artificial, and their selection is significantly influenced by experimental design considerations, rendering them somewhat biased. Admittedly, there's no a priori reason to assume that these formats will seamlessly align. However, the overarching objective of neural decoding paradigms and the corresponding experimental pipelines is directed towards the eventual convergence of information about and within the system. We don't find this observation compelling enough to necessitate an instrumentalist stance, suggesting that these procedures merely constitute approximations or interpretations of the “genuine” informational currency.

The notable achievements in making systematic predictions about neural firing patterns, reliant on the content of experiences and facilitated by learned similarity relations delineated through RSA-type methods (as evidenced in Thornton & Mitchell, 2017), bolster the argument for the causal efficacy of semantic relations. While it's justified to acknowledge that the current methods are still relatively crude, particularly given their operation on a level considerably coarser than the actual intricate neural message exchange (considering that a single voxel contains hundreds of thousands of neurons along with four times as many glial cells), their demonstrable predictive capabilities should be seen as a testament to their potential rather than a hindrance. If these provisional representations, which are already performing impressively, are put to such effective use, it's reasonable to

anticipate that a more fine-grained approach could yield even more refined mappings. In essence, ruling out the eventual alignment of these currently approximating formats would be premature. On the contrary, the synergy between the theoretical arguments we present, particularly the semantic account of similarity, in conjunction with their practical success, should strongly indicate that a growing convergence is indeed feasible.

Even in the face of empirical data that does not perfectly align with the underlying mechanisms, it remains evident that RSA findings, unless entirely fabricated, must inherently signify something that corresponds to real patterns in the world (Dennett, 1991). Its predictive efficacy is indicative of a systematic discovery, one that reveals real patterns, albeit partially veiled. RSA offers insight into the underlying constructs, shedding light on real patterns in the process.

This situation closely mirrors the limitations of a measurement device with inadequate resolution. Just as microscopic events can be captured structurally with a naked eye without revealing every minute detail, RSA results, although less detailed and somewhat abstract, still reflect tangible real patterns in the domain under investigation. This places the onus on critics to demonstrate that these observed correlations are mere artifacts of RSA analytical methods and devoid of substantive significance, rather than real patterns.

These structural relationships, such as semantic clustering, inherently present in the data, can only be elucidated by postulating their connection to the underlying mechanisms. These real patterns can be explained solely by reference to the operational dynamics of these mechanisms.

Psychological Models: However, are the theoretical arguments and predictive capabilities alone sufficient to uphold the position we advocate in this chapter? A second counterargument suggests that what's missing is an additional criterion for

aligning cognitive architecture (or geometry) with the structure of the world. This intriguing line of thought has recently been explored by Bowers and his colleagues (2022) in their extensive discussion on the suitability of deep neural networks as models for biological vision. According to these authors, any successful decoding scheme ultimately requires a psychological model to demonstrate its explanatory power.

This necessity arises from the fact that successful explanations must not only showcase their effectiveness in terms of "results" but also elucidate the reasons behind why a particular result (or a set of results) has been achieved. In other words, we seek evidence that decoders make their operational decisions based on the same principles as the target cognitive systems. Bowers and his colleagues present a multitude of examples that illustrate that this is not consistently the case for deep neural networks (DNNs) thus far. While delving into the specifics of these examples is beyond the scope of our current discussion, some instances involve examining whether part-whole relationships are faithfully represented (which is not always the case) or if the same aspects are prioritized in categorization tasks (where humans predominantly rely on shapes while artificial systems often favor textures).

But what if the patterns purported by RSA-type methods are merely epiphenomenal and do not genuinely mirror patterns found in the environment and structurally represented within a cognitive system? Consider that, in the case of DNNs, it has been demonstrated that the rationales behind input-output decisions are notably distinct from the cognitive processes humans employ. In such scenarios, can we confidently assert that structural representations are indeed at play? Particularly when, in some cases, the algorithms used by artificial networks may be less sophisticated than those postulated by psychological models (which Bowers and colleagues have aptly labeled as "shortcuts"). While it's true that human

cognitive mechanisms also incorporate shortcuts, we can contemplate these divergent decision-making problems in a more agnostic light—as issues of confounders. If a different variable, not related to semantic properties or their content, better explains the behavior of a particular system tackling the same task, is it not a foreboding sign for a theory positing the mirroring of semantic properties as the key to these systems' success?

In essence, we regard both of these arguments as well-founded and valid. Nonetheless, neither of them proclaims that the endeavor of neural decoding based on matching semantic information is fundamentally flawed or conceptually impossible. Neither Ritchie and his colleagues nor Bowers and his co-authors perceive an intrinsic reason for such an approach to fail. Their plea, in essence, is a plea for refining the protocols, devising more stringent tests for isolating confounding variables, and augmenting the explanatory toolkit with psychological models.

We concur that RSA-type analyses and other semantic matching procedures remain imperfect. Nonetheless, they are on the right trajectory, particularly when juxtaposed with their predecessors in decoding methods, such as classical multivariate pattern analysis (MVPA) (as introduced by Haxby et al., 2001). Previous methods were limited to deploying classifiers without delving into the postulation of internal relations between representational geometries. Partially because of this, previous decoding techniques were criticized for their inability to establish determinate content—the mappings appeared to be too arbitrary without a robust interpretation of the internal structure.

For instance, the practice of isolating semantic content within the brain through principal component analysis (PCA) has recently faced a challenge posed by Burnston (2021). According to his perspective, PCA (or similar methods such as

independent component analysis—ICA) does not shed light on the actual “chunks” of cognitive processing within the brain—those stable vessels responsible for housing content. In essence, Burnston argues that components unearthed through statistical analyses do not genuinely correspond to any psychologically tangible primitives. He regards PCA as a fundamentally flawed endeavor due to the brain's multistability in signaling—although we shall not delve into this particular issue here due to space constraints (for a response to Burnston, see also Kinsey, 2022). These concerns underscore the inadequacy of mere decoding without a foundation in psychological models.

Nevertheless, as previously highlighted by Ritchie and his colleagues, RSA-type methods effectively offer the missing psychological model to complement the mere generation of predictions about the content of experiences from neural firing patterns (commonly referred to as “reverse inference”) using statistical procedures alone. In contrast to principal components, which often lack clear, interpretable content, semantic categories alleviate these concerns. Therefore, in our framework, internal informational relations are further restricted, as they are governed by semantic and taxonomic properties that structure the representational space. To summarize, representational mechanisms grounded in structural similarity are not absent from psychological models; instead, they stand as primary candidates for organizing these types of models.

Epiphenomenal Correlations: As for concerns regarding the epiphenomenal nature of patterns uncovered by RSA-type methods, the ultimate test lies in the practical outcomes, determining whether the spaces charted by Kriegeskorte and his colleagues indeed represent definitive maps of perceptual spaces within the brain. We've already discussed apprehensions about granularity and format, which reinforce skepticism that current categorization is not entirely straightforward. We've also acknowledged that current categorizations only encompass simple

relations and a limited set of labels, such as animate/inanimate, humans/non-humans, or body parts. Clearly, this mapping falls significantly short of the specialization and complexity characterizing real cognitive systems. Nonetheless, these instances are better regarded as proof-of-concept or preliminary examples—at best, early endeavors aimed at eventually aligning external mappings of cognitive geometries with the actual internal information processing. In this sense, current RSA-based results can perhaps be perceived as somewhat artifactually or epiphenomenally influenced, but not in the stringent sense of being illusory or incorrect. It is an empirical matter to decisively determine the criteria or categorizations, including cluster boundaries, their number, distances, and the precise details of the representational geometries. Once again, our framework remains impartial concerning specific mappings; their empirical establishment is a prerequisite. We argue for the essential existence of such structures as causally active, mechanistically significant, and informationally grounded.

Importantly, the causal relevance of vehicles can be demonstrated by studying evaluation mechanisms. While it's challenging to experimentally study them, research into evaluation mechanisms related to prediction error and metacognition has shown promise (for more detail, see Bielecka & Miłkowski, 2020; Buckner, 2022). In the context of RSA, some preliminary research on understanding evaluation mechanisms involves testing variation patterns, which subsequently constrain the range of potential ascriptions. Moreover, only specific types of content can effectively explain the causal roles of particular signals, with no clearer manifestation of this limited arbitrariness than in the case of semantic clustering. Contents turn out to be causally efficacious as mechanistic components because they account for the consistent emergence of these semantically organized structures.

The theory of representational mechanisms, therefore, offers a potent heuristic: when you suspect the presence of a representation within a cognitive system, direct your attention to the evaluation mechanisms and study the subsequent processes. To put it differently, what renders RSA an incomplete account of representation is its limited capacity to systematically address this question. Nevertheless, when coupled with the theory advocated here, RSA can be employed to propose additional experimental studies.

No novelty: Critics might argue that our Correspondence Network Framework merely applies existing ideas from structural representation theories (e.g., Shea, 2007) with different terminology. However, this objection misses several crucial innovations. First, while Shea and others acknowledge the importance of type-token relations, they do not systematically address how multiple classification schemes operate concurrently over the same vehicles. Second, our framework uniquely explains how cross-channel evaluation enables content-sensitive processing through network-level operations. Third, we provide a formal account using infocorrespondence to explain taxonomic organization in neural representational spaces. Fourth, our framework directly addresses anti-representationalist critiques by showing how content becomes causally efficacious through network architectures. These features collectively constitute a novel theoretical synthesis that extends beyond previous accounts.

6. Conclusion

This chapter has introduced the Correspondence Network Framework, integrating correspondence-based semantic information and representational mechanisms in a novel theoretical synthesis. While structural similarity has been recognized as important to representation (e.g., Shea, 2007; Gładziejewski, 2015a; Shea, 2018), our

framework uniquely explains how networks of correspondences across multiple channels enable representational mechanisms to systematically process content-sensitive information.

Our framework extends beyond previous accounts in several ways: it explains how neural vehicles can simultaneously participate in multiple classification schemes; it provides a principled explanation for why neural representational spaces consistently organize along semantic dimensions and, as such, avoids troubles for traditional explanations based on arbitrary coding schemes. Additionally, it directly addresses anti-representationalist critiques by showing how and why content has to become causally efficacious (and indeed does, as documented by empirical studies) through network-level operations that detect cross-channel consistency.

Examining RSA through our framework revealed how this methodology implicitly relies on taxonomic organization emerging from correspondence networks. This demonstration highlighted both the utility of our approach for understanding representation in practice and its potential for guiding future empirical research. By offering a formal account of how multiple classification systems can operate over the same vehicles, our framework provides a powerful experimental heuristic for studying cognitive mechanisms.

Our account explains the non-arbitrary binding of targets to contents through correspondence networks that preserve taxonomic structure across processing channels. By addressing both how representational mechanisms process information and why researchers' inferences about these systems are informative, the Correspondence Network Framework brings coherence to disparate notions of semantic information by revealing their grounding in networks of structural similarity.

Overall, this approach provides a powerful, pluralistic framework for conceptualizing representational mechanisms across the cognitive sciences, bridging philosophical formalism with empirical techniques, offering new analytic resources for investigating the foundations of mental representation.

Chapter 8: On Generating Information

1. Introduction

We have argued that cognitive systems compress environmental information into efficient formats (Chapter 6) and validate content through structural correspondence (Chapter 7). But efficiency and matching are not ends in themselves; they are the prerequisites for the mind's most potent capacity: generativity. The ability to produce novel, plausible information—to imagine, plan, and extrapolate beyond the training set—is the ultimate functional test of the Neo-IPP[74]. If our account of structural representation is correct, it must explain how a physical system can generate genuine novelty without resorting to magic. This chapter defines generativity by distinguishing it from mere interpolation, arguing that the generative power of the mind is the direct computational yield of the optimized compression mechanisms defended in the previous chapters.

---

[74] The puzzle of how finite input yields novel, generalized output is not unique to the Neo-IPP; it re-litigates the classical problem of induction tackled by Plato (see: *Meno*), the British Empiricists (via Locke and Hume's transformational abstraction), and more recently by Buckner (2018; 2022). In perception, this lineage extends through the Helmholtzian tradition of 'unconscious inference' (Hatfield, 2002), which positioned perception as a constructive, quasi-generative act, a view later radicalized by the 'New Look' psychology (Bruner, 1957) and its modern descendants in predictive processing. Today, this ancient problem has been re-engineered into the defining challenge of the 'generative AI' boom: how to build systems that do not merely parrot training data but genuinely extrapolate. The goal of this chapter is simply to provide a more stringent definition—one that might well revoke the 'generative' credentials of this entire new wave of AII. No pressure then."

2. Generation vs. faux-generation (fixing the explananda)

The first step for moving towards finding a satisfactory definition of generativity ought to establish the target. For some may say the problem is trivial, as generativity is inherently a very vast phenomenon, which occurs wherever new information is created, and as such should be defined very loosely. Why then not simply say that the only criterion for generativity is producing a new piece of information? This would seem to offer a clear and simple solution to the problem.

Such a resolution might be elegant but would not be very useful, though. It is not difficult to see why. If we consider two transmitters of information, one that spits out truly random numbers and the other one that repeatedly sends only zeroes, we are inclined to say that the former is more generative than the latter. Of course, the difference in this rather silly example is immediately understandable with only basic Shannonian theory - the measure of entropy is what licenses us to posit the former coding scheme is maximally informative, whereas repetitive transmission is negligibly informational. What is important for us now is just observing that merely churning new strings that eat up (memory, channel) space is prima facie different from truly new information being produced. The simple solution trivializes the problem, rendering it uninteresting.

Adding available degrees of freedom (symbols at hand) to our zeroes only sender would not help, either. If we think of neuronal architectures as trading in information, simply adding more neurons does not translate into informational gain. It does endow the system with additional power to potentially hold and process more information, granted. However, neurogenesis itself is not yet the informational brain gain. The same goes for establishing new connections between neurons (or between their parts) and for the strengthening of these. Even so, these are widely postulated as mechanisms of learning (e.g., Kennedy, 2016;

Castello-Waldow et al.,2020) - a paradigmatic case of informational gain, at least intuitively.

Of course, I am not denying here that these mechanisms are not plausibly necessary for true generativity. So far, I am merely pointing out the fact that it's not enough if we want to differentiate between actual and faux generativity.

Again, transforming just what is - in a sense - 'already in there' does not qualify as generative, either. Otherwise, the bar is on the ground. Every new output would be 'generated.'

It is not particularly difficult to come up with a slew of routine computational processes that we would be hard-pressed to categorize as generative, even though they produce 'new' input. Copying is the most obvious one. If I have a file (or a picture) and I make a backup or xerox the physical object, a new token is generated. However, regardless of whether it is identical (as in the case of a computer file, in principle) or only ever slightly degraded by noise, there is information transfer (as Dretske would agree), surely, but there is no real generativity. And we can think of more complex procedures that would not clear the bar either. Translation (in the geometric sense) is a transform that ends with a different (shifted) output on a rudimentary level. So do lossy (irreversible) transforms, which perhaps helps get the point across. The final product seems to be already entailed in the input. Importantly, there does not seem to be a major difference whether we consider something like downsizing to—for - for instance - data merging. In both cases, we end up with less information than we 'had' before the transform, and - again - we merely transformed, not created any new data.

The cases, however, can get murkier. And indeed these murkier waters we want to probe to look for a satisfactory place for a delineation. Let's then consider procedural generation (proc-gen). It is a computational technique explicitly

designed to produce 'new' (potentially arbitrarily many) inputs based on a much smaller set of generation rules. It can provide an abundance of riches in terms of a basic toolbox, as any person who has sacrificed thousands of hours to a fiendishly and deceptively roguelike game would be happy to confirm. However, proc-gen is famously rigid and gimmicky. Mentioned gamers were often unhappy when the scenarios became too scripted and predictable, pointing out the fact that procedural generation is but a trick designed to mimic real creativity (Colton & Wiggins 2012) or novelty - albeit a very clever one.

How are these faux-generative processes different from truly generative ones we are trying to pin down? How does new information emerge that is not already entailed in the available or accessible data? We need a contentful understanding of generativity. However, this upgraded understanding of generativity has to avoid the pitfalls of overmentalization. We want naturalistic explanations with biological and artificial systems to serve as proof-of-concepts or models of the real generativity being in principle achievable without resorting to human performance as the only source of novel informational input, as certain phenomenological trends within philosophy tended to do, borrowing the idea of 'generativity' from Husserl's philosophy (Steinbock, 1995).

If we are serious about looking for true generativity under the personal-level description, then perhaps there are already concepts and tools in information engineering already in place that could help understand where new information could come from. Information theory is an obvious candidate for this type of insight.

Usually, in the literature—both technical and philosophical, building on information theory—more focus was devoted to the notion of informational loss. Famously,

Dennett (Dennett, 1991[75]) argued for the existence of real patterns on the principles of algorithmic information theory (discussed in more detail section 4.5 of ch. 1), where the former are defined over the limits of lossless compression. Similarly, in practical application, the veridical and efficient transmission of signal has more to do with the loss of information (and ways to limit, decrease, or circumvent it) than information net gain. The existence of a source is usually assumed as an entry point for problem-solving (even if the properties of the source are not known and the problem is indeed tackling this uncertainty). Anyway, the etiology of information is rarely the point of interest. Nevertheless, information-theory is equipped with a notion of 'information gain' (Cover & Thomas, 2006), and it is hardly a niche concept. It is, for instance, extensively used in constructing decision trees in data science. Information gain is a measure of the quality of attributes that could be used for parsing a given dataset. A feature that maximizes information gain is treated as a strong candidate for a good decision criterion.

Unfortunately, the 'informativeness' here is purely a technical one. Information gain in this context equals just the Kullback-Leibler divergence of a particular variable. It is a measure of how much information about a given signal could be obtained by learning the value of another variable. Importantly, then, information gain is a relational property stemming from the differences between a true and estimated (used) distribution. The word 'gain' comes from the fact that substituting the distribution closer to the real one yields an increase of informativeness from the one less similar to the true distribution (the KL measure is hence the quantification of this increase; Burham & Anderson, 2002). Information gain does not tell us much about how to differentiate 'real' generativity from an ostensible one.

[75] For an interesting recent revival, see: Millhouse, 2022

3. Generativity in neural nets

A second obvious place to search for ready answers is 'generative models.' The word had become commonplace in the machine learning community already in the 1990s (Schmidhuber, 1990), but it has really exploded onto the scene with the emergence of deep neural networks (DNNs) and then transformers. Philosophical literature has not turned a blind eye to this fashionable concept. Especially, the predictive processing framework has adopted the idea of 'generative models' as a core concept in their project of rewriting the explanatory project in philosophy of cognition, starting with Andy Clark's work on Rao and Ballard's model (Clark, 2016; Rao and Ballard, 1999).

Generative models are usually defined as statistical models of the process of how sensory data are generated from a set of hidden causes. They are functionally different from more classical discriminative models. While the goal of the latter is to learn the boundary of the classes, generative models are aimed at learning true distributions of the data to produce new samples based on the training data.

Technically speaking, the beginnings of generative models can be traced even back to the 1950s with the emergence of Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). Yet, these methods, in part due to technical (computational) limitations, did not cause a major shift in artificial intelligent systems design in a way that we associate with modern generative systems. A partial breakthrough came in the 1980s with the discovery of recurrent networks (RNNs; Hopfield, 1982), which allowed for more convincing and variational generation of previously unseen inputs. An even more impressive result was achieved with the introduction of LTSM (Long Short-Term Memory; Hochreiter & Schmidhuber 1991), which significantly improved the performance of the models with longer strings.

However, up to the 2010s, the models were mostly deployed for analyzing numerical data, sometimes texts. They struggled with other modes of data, including ones that were most sought after from a market perspective - like speech or images. That has largely changed with the introduction of variational autoencoders in 2013 by Kingma and Welling (2013). “VAEs opened the floodgates to deep generative modeling by making models easier to scale [...] Much of what we think of today as generative AI started here.” (Akash Srivastava; quoted in (Martineau, 2023)).

Autoencoders were revolutionary, partly because they consisted of two parts, working somewhat in opposite directions. The first one is an encoder, which is tasked with mapping unlabeled data into a compressed representation. These representations are stored in the space called the latent space, which can be conceptualized as an internal map of the model created on the basis of the training input. Regular autoencoders keep single entries, that is, values of particular variables. Variational autoencoders’ improvement was to store not values, but entire distributions within the latent space. That allows for the second part of the pipeline to achieve much more interesting properties.

That second part is the decoder, which intuitively does the encoder’s work in reverse. They invert the mapping, linking the latent space representations with the original type of input (now as output). For ‘regular’ autoencoders, that meant reconstructing the original message, and that’s why they were used for denoising, such as reconstructing corrupted files, detecting errors in speech, or deblurring images. However, the aforementioned critical feature of VAEs that store distributions, not points, enabled them to sample from this space. The effect was that the outputs were alike to the original message, but not identical - tokens of the same type, but not copied or merely decompressed. Whereas more traditional coding schemes achieved transmission and processing of original data, VAEs paved the way for outputting variations of the original data.

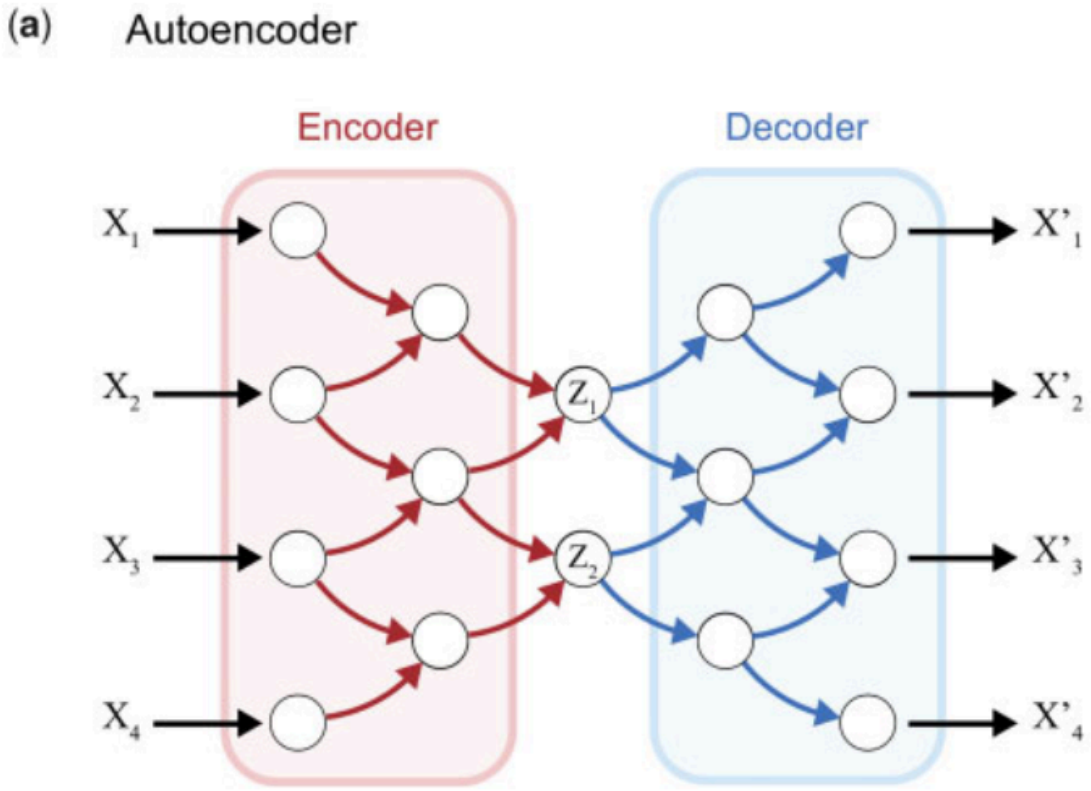

Figure 12: In autoencoders, the encoder part (shown in red) compresses sensory information to compact representations in a latent space. This representation is decoded into sensory data format. The decoder (shown in blue) can be used for counterfactual information generation using a seed chosen from the latent space. The variables z1 and z2 represent the latent variables. [after Kanai et al.,2019]

On the face of it, then, generative process happening in the decoder part of the autoencoder is realized through the decompression from the latent space, which is the seat of 'representations' (in the most technical, machine learning sense of the word, where it is satisfied merely by the stand-in relation) for the model. But it's not a regular decompression, as it is merely an expansion of what's already there.

To understand the difference, think of zip files, then compare them to the technique called upscaling.

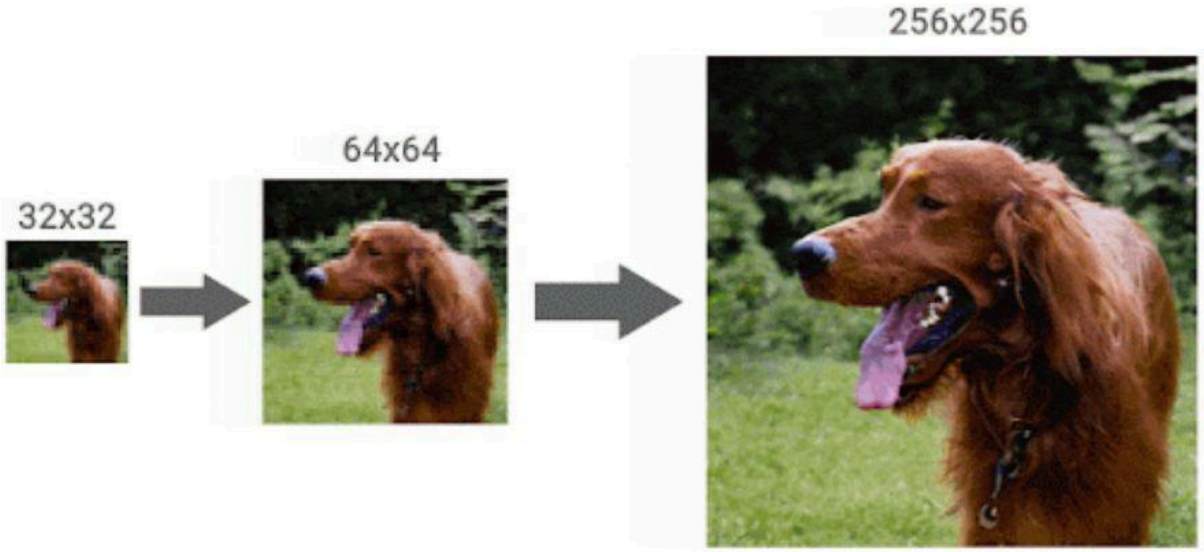

Figure 13: Example of image upscaling. The same image of increasing resolution is inferentially upgraded in order to preserve the high fidelity to details even on a bigger scale. The feat is usually not possible using 'direct' image processing.

Traditional compression, while an art in itself, is relatively limited. For starters, the lossy compression is not reversible. That is immediately clear to anybody who tried to magnify a JPEG and was disappointed to find out that the bigger the picture, the more the quality of the detail goes astray. That is because there is simply not enough information in the file to decide which pixel values should be taken between the known ones - if we extend the image, the empty spaces appear between them, and decision gimmicks implemented in the JPEG protocol lead to the impression of blurriness and deterioration of details[76].

[76] To be precise about the reality of these 'gimmicks,' what happens when we extend the image is that the spaces between pixels must be filled, and the interpolation algorithms (such as bicubic or bilinear filtering) implemented in standard protocols inevitably smooth out high-frequency details, leading to the impression of blurriness and the exposure of compression artifacts

This problem is impossible to overcome without resorting to the background knowledge of images in general (at least to ones we tend to call 'natural,' according to the Natural Scene Statistics camp; certain power laws (autocorrelation, power spectra) apply to images that are, well, worldly). Operating solely on the file itself can only provide limited returns. The utilization of these constants enabled the development of protocols that allow for reliable and noticeable differences in the ability of image storage protocols to 'fill in' the unknown details (values) of pixels that are not directly stored within a file. Using histograms of known distributions, methodologies developed by Bill Geisler and colleagues (Natural Scene Statistics researchers) enabled them to 'upscale' images with remarkably higher levels of precision. But these protocols are not mere decompressions; they utilize known dependencies and regularities in the signal to make informed inferences about the missing information and fill in the blanks with this content.

Coming back to the autoencoders. When one studies Figure 15 (below), it quickly becomes apparent why these sorts of advances in artificial models generate such excitement among proponents of predictive processing, Bayesianism, and advocates of generative models in cognition in general. After all, the encoder-decoder pipeline resembles, quite strikingly, the picture presented by these constructivist approaches in cognitive science. For a PP proponent, cognitive architectures hinge on top-down inferences. In more radical versions, where all perception is prediction, opportunistic guesswork (Clark, 2015), or 'controlled hallucination' (Clowes, 1971[77]), 'hypotheses' (in their very particular understanding of the term) or hyperparameters are responsible for generating predictions for the lower level. Bottom-up processing is then readily modeled by the encoder in the VAE pipeline, whereas top-down processing is modeled by the decoder. It is

[77] Even though it was coined by Clowes, the phrase became a bit of a catchphrase for predictive processing movement, thanks to its popularization by Anil Seth in both scientifi publications (2021), but also a plethora of media appearances (to be clear: Seth openly credits Clowes with inventing this witty phrase).

instructive to compare the previous diagram with the PP version of the encoding-decoding duo:

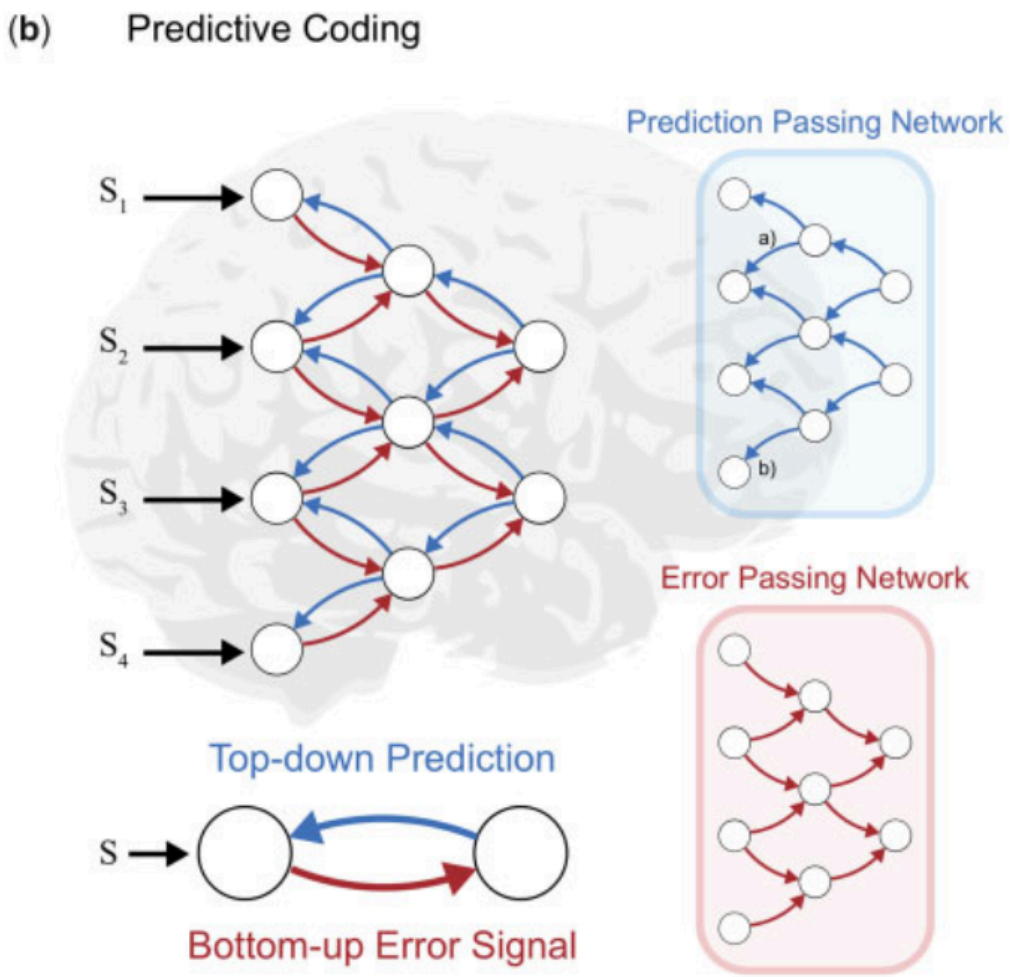

Figure 14: Top-down and bottom-up information transfer

The colors hold their meaning across both diagrams. We can see that the encoding part's counterpart (red) is the 'error-passing network' in PP lingo, which is roughly the 'traditional' bottom-up direct perceptual input being propagated up the network. It is responsible for fixing the latent space through a learning process, which results in the formation of the internal model that tries to approximate the

structure of the environment (totality of inputs). Hence, the formation of the internal model is nothing more than the process of representation learning.

On the other hand, the 'blue' part of the schema is the reverse process - the decoding/generating side. On PP reading, this process is equal to producing sensory events based on information stored at the hierarchically higher levels of the system. Respectively, for VAEs this amounts to variational decompression from the latent space onto the original input space after sampling a seed from the latent space.

Proponents of VAE-PP marriage put it succinctly: 'We argue that information generation can be seen as production of sensory representations using internal models. This is achieved by reversing the process of representation learning, which is projection from sensory inputs to internal models.' (Kanai et al.,2019:4)

Also the semantic pointer research program (Thagard & Stewart, 2014; Eliasmith, 2013) adheres to similar principles when discussing how concepts work within their framework. For instance, Blouw et al.,(2015: 1132-3) write:

'In its most basic form, a semantic pointer can be thought of as a compressed representation that captures summary information about a particular domain. Typically, such representations derive from perceptual inputs. An image of an object in one's visual field, for instance, will initially be encoded as a pattern of activity in a very large population of neurons. Through transformations of the sort described above, however, further layers of neural populations produce increasingly abstract statistical summaries of the original visual input [...] Eventually, a highly compressed representation of the input can be produced. The reason compressed representations of this sort are called semantic pointers is because they retain semantic information about the states they represent by virtue of being non-arbitrarily related to these states through the compression process. The reason why the representations are referred to as pointers is because they can

be used to “point to” or regenerate representations at lower levels in the compression network (Hinton & Salakhutdinov, 2006).’

We see here both crucial points of the picture. Concepts are conceived as highly compressed bundles of the original input, generalized, abstracted, and stored within internal models (that is pretty standard even for ‘traditional,’ i.e., purely bottom-up accounts of perception). But they are also functionally responsible for pointing to lower-level addresses in perceptual memory (the metaphor is here very literal; semantic pointers are explicitly inspired by memory addresses in computing machines) in order to regenerate information that they encode for. For this reason, they can guide categorization as they bundle and organize all relevant information pertinent to the task requiring this concept. It is important to note here that on this reading, concepts are unstructured, contrary to, e.g., coherentists (following Sellars (2007)), where concepts are highly interconnected and embedded in inferential relations that constrain them.

A very similar account of conceptual role in generative processing has been offered by Quilty-Dunn’s paper ‘Polysemy and thought: Toward a generative theory of concepts’ (2020), in which it is argued that ‘regular’ pointer architectures (Gallistel & King, 2011) are insufficient for explaining away polysemous phenomena. The crux of the idea is to salvage the compactness of concepts as compositionally friendly and multipurpose tools of thinking without sacrificing the ability to explain how they can still refer to separate chunks of information they refer to (richness). It could suggest the kinship to ‘mental files’ (Recanati, 2012) type of theories, with the difference that the associationist commitments of these theories are rejected.

Interestingly, these needn’t even have to be ’bank'-type scenarios where the same label refers to strictly different bodies of information, here - money vaults and river coasts (among others). Separate - well - banks of information could be subtler and

cater to different functions of the same object, for example. The example Quilty-Dunn gives is doors, which can be summoned in cognition as 'door-qua-apertures' and 'door-qua-barriers,' invoking both different functional roles and separate bodies of information.

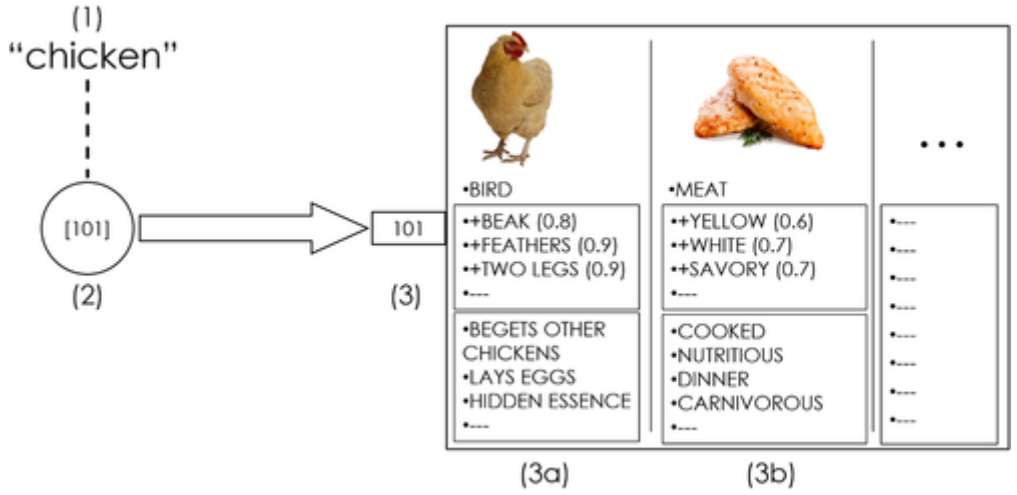

Figure 15: Yet another example of the same label that could point to separate arrays of information stored at the lower level. Chicken-qua-bird is stored at a different memory address than chicken-qua-meat, even though they are served by the same concept, 'chicken.' Chicken is thus a generative pointer insofar as it could generate one or another, contingent on the task at hand.

To borrow terminology, concepts in this framework are not merely activated but deployed. Calling the concept equals modulating the relevant denotation, not simply rigidly referring to it. As far as I am skeptical of whether this move truly enables escaping the troubles of original conceptual atomism of Fodor and followers (e.g., Lawrence & Margolis, 1999), and it does seem to me that Quilty-Dunn's solution just potentially shifts the problem of disambiguation one level down without explaining how endowing concepts with diverse quasi-associative does not infringe on their compositionality; this problem should

not concern us here. What is important for our purposes is to establish that the generative nature of concepts in both semantic pointer architectures and predictive processing strands of theories follows closely the principles postulated in artificial nets engineering, dubbing themselves generative. This suggests that there is some shared understanding of generativity that exceeds mere creation or production of new information. However, it is usually not spelled out explicitly there.

While artificial architectures like VAEs and semantic pointers demonstrate the computational feasibility of generative decompression, they leave open the question of biological implementation. Furthermore, they face the persistent philosophical charge that compressed codes are merely amodal symbols, divorced from the phenomenological richness of sensory experience[78]. To defend the account of generativity proposed here, we must show how the brain physically implements the 'decompression' of abstract vectors into rich simulations without losing their grounding in the sensory cortices. This requires bringing back Phillips' theory and integrating its biophysics of the two-point processor with the control-theoretic and embodied frameworks of offline cognition.

---

[78] This objection conflates the storage format with the generative process, a misunderstanding resolved by integrating the "simulation" theories of Barsalou and Grush with the deep learning empiricism of Buckner. Barsalou (1999) explicitly defines concepts not as static symbols but as dynamic "simulators" that rely on stored "schematic frames" (compressed invariants) to reconstruct rich perceptual experience. Similarly, Grush's (2004) control-theoretic emulator runs offline by processing a compressed "efference copy" to generate a detailed "corollary discharge" (a simulated sensory outcome). Buckner (2023) synthesizes these views, arguing that modern deep neural networks vindicate this empiricist architecture: they demonstrate that "abstract" concepts are simply regions in high-dimensional vector spaces that disentangle complex factors of variation. Buckner specifically critiques the tendency of philosophers to overvalue phenomenological intuitions about the "hardness" of multimodal integration, noting that for modern engineering systems, integrating diverse modalities is now a routine operation performed by mapping inputs onto a shared, lower-dimensional manifold. In this latent space, "abstract" does not mean "amodal" in the traditional sense of being arbitrary; it means being positioned on a manifold that captures deep structural invariants. Thus, the "compressed code" is not a barren amodal symbol but a generative kernel—a vector that, when unpacked by the right architecture, yields the full sensorimotor simulation.

4. Interpolation / Extrapolation boundary

Luckily, there is an interesting conceptual debate going on in the machine learning community that tentatively could help provide a more clear-cut way to pry apart genuinely generative from faux-generative processes. This debate hinges on finding the boundary between interpolation and extrapolation.

If we go back to the upscaling example from part 3, we recall that what enabled the 'supercharging' of images was the inference-backed ability to 'fill in' the in-between pixels. The idea is simple: when we want to guess the value of the pixel in question (which is the blank spot after extending the image), we take what we know—the value of the pixel preceding and the value following—and try to estimate what the middle value should be. We do that based on prior knowledge about the internal relations between pixels in natural images.

Now we can generalize this thinking about inferring the 'middle' values between known data points and take it to apply to different modalities of data and also higher-dimensional datasets. If we consider learning algorithms, then the known data points in the previous examples would correspond to the points in the training regime. They are the known values, the already-seen input. Now, what we want from our systems (predictive ones, at least) is to be able to - well - generalize reliably from these known sets of points onto the unknown.

Now, when we think of known data points, we can delineate a boundary around them - its geometric closure. When we consider the simplest 2-dimensional case, where points lie on the X-Y axis, the most common type of that could be the line of linear regression. Now, these generalizations can occur in two ways.

The first is called the interpolation regime, and it entails the types of cases we discussed previously - when the value to be predicted lies between known data points. When this condition is not met and the value in question lies outside the

geometric closure of known data, then one enters the extrapolation regime. This difference is very straightforward on the face of it, and as such is routinely taught in virtually every major data science course. The diagram below illustrates the difference in the simplest terms:

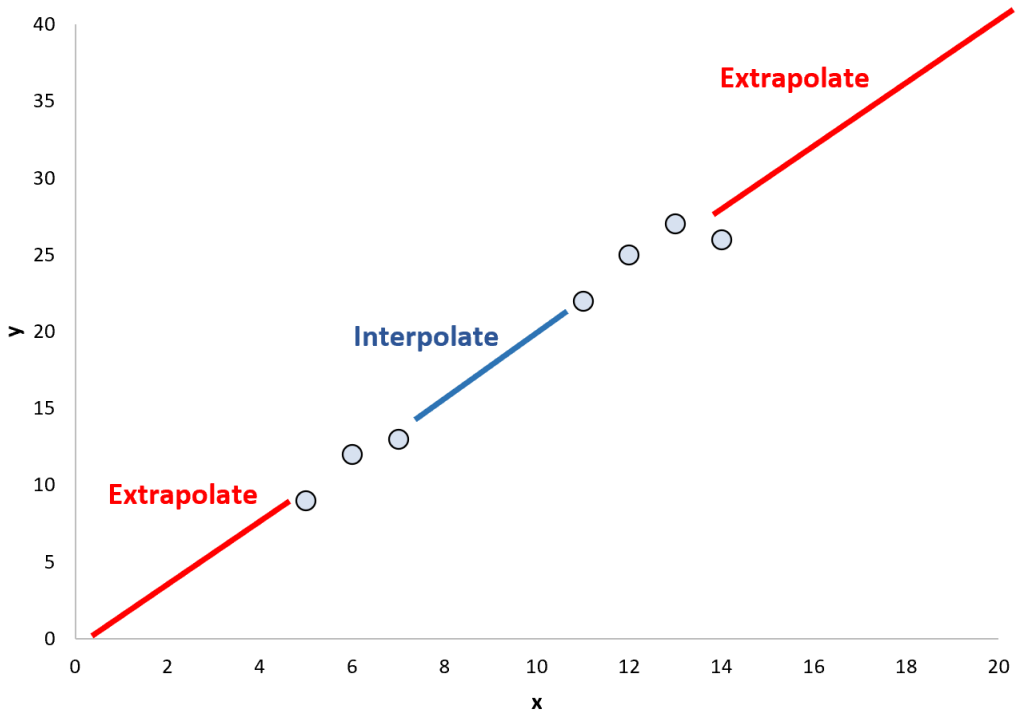

Figure 16: Difference between the interpolation (blue) and extrapolation (red) regimes. The former is reserved for inferences between known data points (blue dots) on the regression line, whereas the domain of the latter lies outside of it.

As said, this hardly seems controversial, even in higher-dimensional spaces. The interpretation is clear. Interpolation is the prediction of data points that lie on the manifold of the training data, while extrapolation is the prediction of data points that lie outside that manifold

We have talked mostly about interpolation so far, but the extrapolative abilities are the ones most sought after in information technology and engineering. The reason

for that is simple - extrapolation is plausibly connected to generalization. In fact, this has been explicitly discussed after initial major successes of ANNs in the 2010s. While very impressive in various use cases, critics (and some proponents; e.g., Chollet (2023) were worried about the abilities for more flexible, generalized behavior of these models. According to them, what the models only do is essentially only 'curve-fitting' (Marcus, 2018)—getting - getting excellent at assessing the distributions of variables in the data, but only within its training regimes. Just to be clear, this line of reasoning was not discovered only after the carnival of deep nets in the 2010s, but much earlier. For example, Haley and Solloway (1992) as well as Barnard and Wessels (1992) warned as far back as the early 1990s of the limitations of systems that do not extrapolate well beyond their training data.

And the real challenge - the critique went - is to escape outside, that is, to truly predict, not only fill in. That's why extrapolation was widely agreed upon to be an adequate operationalization of the ability of models to generalize and abstract well. It also seems to capture well the idea important for our purpose of pinning down genuine generativity, as it appears to give a precise mathematical formulation to the difference between true generation and mere unfurling or expansions of the already known information.

However, recent literature suggests that the picture is trickier, which spells potential trouble for us if we want to delegate explanatory work to these mathematical concepts. That is because the intuitions, built upon lower-dimensional cases, seem to fall apart for higher-dimensional spaces. This is exacerbated by the fact that sophisticated cognition surely has to tackle high-dimensional input spaces well to be efficient without giving up on the richness of the environment.

But why would things fall apart for high-dim spaces? Shouldn't the same principles of delineating along the closure of training data hold regardless of In principle, yes, but the problem is subtler and more practical. Balestriero et al.,(2021), in a recent study, demonstrate that for spaces whose dimensionality exceeds around one hundred (R > 100), true interpolation is very unlikely to be achieved in practice. One way to think of it is via the aforementioned curse of dimensionality - in highly complex spaces, points are significantly more scattered. Any new given point to be predicted is very unlikely to lie inside the closure of data.

There is perhaps a more intuitive way to think about it, somewhat Monte Carlo style. If we think of our manifold defined by known data as a multidimensional target and we try to throw darts at it, the entire space that we shoot at is so vastly bigger than the target it is very unlikely we will hit anywhere outside the target landing zone. This unlikelihood explodes with the number of dimensions, which explains why elegant and seemingly reasonable intuitions from simpler manifolds seem to break apart when things get fairly large.

This is a significant difficulty, regardless of its mathematical roots. It is in reality a practical obstacle more than an in-principle one. It’s not that it’s impossible to interpolate in high-dim, just very unlikely. However, the fundamental problems of reliability in out-of-distribution generalization remain acute, as demonstrated by Balestriero. After all, we are not eventually concerned with mathematical properties of interpolation or extrapolation. We are concerned with their potential explanatory work. And here even more philosophical questions surface. As usual, these stem from the fact that math is not the territory, and things can get tricky when we want to find correspondences between formalisms and their realizations.

First, it is not obvious which type of geometric closure should be used for setting up the boundary between two regimes. Is it the linear hull of the training data? Is it its

convex hull (as Balestriero et al.,claim)? Or is it some other geometric closure of a set (Milliere, personal communication)? We can call it the ‘boundary problem.’.

Yet, there is a further conundrum: when we consider neural nets, in what space should we be looking for that manifold? Should we look in the input space? Or should we look in the neural space? And if it's the latter, which layer of the network should we be looking at (ibidem)? This could be dubbed the ‘manifold problem.’

Bonnasse-Gahot (2022) takes both the boundary problem and the manifold problem in the attempt to criticize Balestriero and colleagues's results. He zooms on the penultimate layer in a neural network, which is the last hidden layer (one before the categorization decision). He utilizes an autoencoder to uncover what he calls ‘the intrinsic space’ of the layer, that is, the space that underlies the neural activity at this level. According to him, the dimensionality of this space is actually low. As a result, test data on this level lies well within the geometric closure defined by Balestriero (the convex hull of training data), thus satisfying the conditions of interpolation.

Interestingly, similar types of results have been reported in neuroscience literature, trying to unearth the mathematical descriptions of spaces underlying neural activities on certain levels of processing. These studies look for counterparts of what Bonasse-Gahot calls ‘intrinsic spaces,’ that is, latent ‘representations’ of significantly reduced dimensionality (Archer et al., 2014; Gallego et al., 2017; Jazayeri and Ostojic, 2021; Chung and Abbott, 2021)

In conjunction, it suggests that by refocusing the search, one arrives at a different conclusion about the interpolation/extrapolation debate. In fact, lower dimensionality is positively correlated with the performance of the model. Such interpretation goes directly against the worries about the practical impossibility of interpolating, since it undermines the assumption of the fatal ramifications of

dimensional explosion of input data. If one focuses on the hidden layer, one discovers that the effective dimensionality of the representation is sufficiently low for the test data to fall within the convex hull of the training set, thereby rendering interpolation mathematically tractable (Bonasse-Gahot, 2022).

This has two potentially important consequences. First, it demonstrates that by probing the manifold problem, we can not only make the assumptions relying on mathematical basis for delineation more explicit, but also help diffuse some worries. Secondly, however, it undermines the solidity of the boundary problem. For, if Bonnasse-Gahot is right (and the discussion is still ongoing), his results, while alleviating the manifold problem worries, seem to suggest that the interpolation/extrapolation difference is less important than we (following both the consensus and Balestriero) suspected.

The rationale for that is the fact that what turns out to be more positively correlated with good generalization was not the inclusion in the convex hull, but the proximity to the training data. In other words, statistical measures of distance within the test space are more predictive of performance accuracy. Locality/globality (Chollet, 2021. ch. 5) is thus a more important feature than the extrapolation/interpolation distinction, as defined by the geometric closure. For a point lying outside the hull but closer to the training data is on average (at least for some networks) more positively correlated with accurate prediction than the point that is scored higher by distance, but is entailed by the 'interpolation area' (i.e., is located inside the boundary)

Altogether, the proximity principle, not the closure principle, does sound intuitive. We are in general more inclined to ascribe more credence to predictions closer to the already-known cases (Bonasse-Gahot, 2022). Expressing it through the terms of statistical distance bolsters the intuition. However, there may be some deeper

philosophical conclusions here, too. Tentatively, it demonstrates how it can be helpful to make background assumptions explicit when making claims about the interpolation/extrapolation distinction, that is, making it clear that it is conditional on some specific mathematical definition of the delineating constraint (boundary problem) and the search of its instantiations (manifold problem). It indicates that mere mathematical arguments are insufficient for answering problems such as defining genuine generativity, and, as such, they cannot be treated in isolation from broader theoretical considerations. If we agree and concede that resorting to pure mathematics cannot salvage a philosophical project (again), we are ready to plunge again into assessing candidate cases of generative processes.

5. Computational photography as the method of cases

We have so far touched upon the simple cases, such as copying, and slightly more advanced ones, like upscaling. Also, sophisticated, high-level, and famously data-hungry computational pipelines such as generative neural networks were considered. In some way, we have probed candidates for generative processes from both the bottom and the top of the ladder. Now, having conceded that purely mathematical resolutions are insufficient for our purposes, I want to revert to discussing cases, but now aiming for more controversial, middle-ground cases in the hope that we may stumble upon the level of generativity where the difference between faux-generativity and genuine generativity could be plausibly established.

For that purpose, it could be instructive to refer to computational photography as a sandbox to create thought experiments to adjudicate cases through philosophical intuitions. Then we can move towards setting up the relevant definition that could account for the differences between said cases.

Modern computational photography is an interesting place to look for these cases for a host of reasons. First, it is classical in the philosophy of perception to look to photography for analogies and intuition pumps for theorizing about vision, but this is the least important reason. Secondly, shifts from traditional, optical photography towards the computational one mimic well the differences discussed in the beginning of the chapter that concerned the activity and constructivism of image or percept formation; they also relate directly to questions of informational richness of the environment vs. the need for informational completion 'from within.'

This last difference becomes very apparent right from the beginning, when the definition of computational photography is attempted. Mark Levoy (2016) defines it as 'computational imaging techniques that enhance or extend the capabilities of digital photography in which the output is an ordinary photograph, but one that could not have been taken by a traditional camera.' Let's note here two things - 'enhancement' and the counterfactual. Computational photography begins whenever traditional optical photography cannot produce a given output because it needs an additional boost.

And modern photography - even the amateur kind, hidden inside most common smartphones - is filled to the brim with clever tricks and enhancements. Importantly, we can roughly divide them into hardware and software enhancements. Plenoptics would be an example of the former kind. The difference here stems from the fact that many modern photorecording devices are now plenoptical cameras (also known as light field cameras or Lytros,-coming from the first commercially introduced devices). They are capable of capturing the intensity of light in a scene, and also the locations and directions of light rays that end up as parts of the picture. It's achieved through putting additional lenses between the sensor or using a multi-camera setup within one device. This contrasts with conventional cameras, which record only light intensity at various wavelengths. In

an important sense, these cameras can catch a richer informational intake than the traditional cameras do.

The situation is different for software-based enhancements. Another technique called 'stacking' may be instructive here. Contrary to everyday opinions, modern cameras do not start shooting when we press the shutter button. Smartphone cameras pre-shoot even before the decision has been made. Several times in rapid barrages, in fact. These proto-shoots are kept in a small memory buffer that is overwritten every few seconds. Kept entirely away from the eyes of the users, this secret life of phones feature enables them to do the said. For it turns out that using the input from snapshots just directly preceding the actual 'target' picture (one after the shutter-stop decision by the photographer) makes it possible to boost the quality through recombination of features, resolving blurriness and stabilizing the picture. In some sense, then, it is similar to the plenoptics case - the camera simply has additional information in its possession. In another important sense, though, these cases differ significantly. In the stacking case, what happens is the recombination of several images, separated by time differences (albeit very short). Whereas, in the plenoptics case, what happens is a more straightforward enhancement, akin to better illumination or a longer exposition period.

Nonetheless, in neither case does the device contain information that is not available in real time. Nor did it actively utilize any previous, stored information. And that is what we are looking for as a marker of generativity. Even the use of the buffer to 'stack,' or recombine (partly) the pictures, resembles more the compositional photography of Francis Galton (1879) than any genuinely generative practice. Whenever there is the usage of information not stored or recorded directly, then additional information has to come from somewhere, and generativity is automatically one of the prime candidates for that. Let's turn to more sophisticated computational photography techniques, then.

Enter the infamous 'moon case.' Sometime in 2019, a Chinese tech journalist, Wang Yue, on the Zhihu web portal (a Far Eastern counterpart of Quora) published his suspicions about the newly debuted P30 Huawei smartphone, boasting a novel feature called 'the Moon Mode.' The Moon Mode promised to provide high-fidelity, realistic pictures of the lunar surface without any additional utensils, such as magnifying lenses. The promise was supposedly guaranteed and powered by—of course—simply 'AI.' The blogger ran some tests, including blurring the initial input or pointing the camera at surfaces that roughly resemble the lunar surface (such as ping-pong balls illuminated in a certain fashion). The camera was still spitting out high-quality representations of the moon. What's more, however, further investigation revealed that the ready representations of the moon were drawn from a database and simply inserted with some post-processing on the edges to iron out the bluntness of such a maneuver and obfuscate the real process from the user. Although this view has been challenged (Zhang, 2022), arguing that indeed an AI model was leveraged to put details in place instead of simple overlaying (or copy-pasting, to be more blunt), this need not concern us here too much. The uproar about the possibility of texture insert or overlaying suggests that there is an important difference to be drawn between that and 'real' generative completion or between AI enhancement and mere 'superimposed alteration' (Zhang, 2022).

That would not be perhaps too different from the cases discussed so far, but the lunar controversy did not stop. Fast-forward to 2022, when one of Huawei's chief competitors, Samsung, unveiled a new functionality called 'Scene Optimizer' (sic!) in its Galaxy S20 series phones. Accusations resurfaced in an almost identical fashion. User-led investigations and tests led to accusations that the company is covertly adding features and details that are 'not really there,' where one points the camera to. Equally important, the selectivity of the processing raised suspicions. If the system is selective only to objects similar to the moon shield, then the

likelihood of using a sophisticated inference-based engine instead of a simple gimmick seems more plausible and renders the company's marketing statements of advanced AI empowerment as unfounded (if its critics are to believe).

Samsung replied quickly in an attempt to quash the controversy. The arguments they used were interesting and also helpful for understanding the types of processing in cutting-edge computational photography, so I will quote them at length and comment.

> From 10x to 100x zoom, image quality is boosted by powerful Super Resolution AI. At one push of the shutter, up to 20 frames are captured and processed at instantaneous speeds. Advanced AI then evaluates and corrects thousands of fine details to produce detailed images even at high magnification levels. And when shooting at high magnifications, Zoom Lock uses intelligent software to set the image in place so you can shoot with minimal shake.
>
> When taking a photo with the Galaxy S21 cameras and Scene Optimizer is activated, once AI recognizes the object/scene it will work through every step of processing. AI will first start by detecting the scene/image at the preview stage by testing it from an AI model trained on hundreds of thousands images. Once the camera detects and identifies the image as a certain scene, for example, the Moon, then offers a detail enhancing function by reducing blurs and noises. Additionally in low light/high zoom situations, our Super Resolution processing is happening (I.e., multi-frames/multi-exposures are

captured > A reference frame is selected > Alignment and Registration of multi-frame/multi-exposures > Solution Output).[79]

The actual photo will typically be higher quality than the camera preview. This is due to the additional AI-based multi-image processing that occurs as the photo is captured. This idea is described by the manufacturer as follows:

> 'For example, when taking photos of an object, **3 key elements are taken into place. Object detection (when Scene Optimizer is enabled), powerful AI processing and multiple frames enhancement. Each one of these features plays a critical role in order to deliver quality photos**. When combined, these features generate the proper balance between a natural look and detail. **The process starts by identifying an object based on a realistic human eye view, then multi-frame fusion and upscaling adds on by generating a higher level of detail to the subject, finally leveraged by AI deep learning solution it uses contextual assumption to process and piece together all the information** to delivering a high quality result.
>
> No image overlaying or texture effects are applied when taking a photo, because that **would cause similar objects to share the same texture patterns if an object detection were to be confused** by the Scene Optimizer' (As stated by Samsung; cited in Wong, 202g; emphases are mine)

There are a few issues to comment on here, so let's unpack it. At the end we see there is a flat rejection of the ping pong ball types of scenarios that we had seen in the Huawei case. In some way, the critics' worries were accurate - the system is

[79] I rely here on Samsung's consumer-facing description as it sufficiently illustrates the functional logic of the system for the purposes of this philosophical exposition. To be sure, the mechanism is well-grounded in computer vision literature; summoning a slew of papers that lie under the bonnet of this (granted!) marketing formulation of 'AI processing' includes, for instance, Wronski et al. (2019) on multi-frame super-resolution and Zhou et al. (2014) on deep scene recognition. I will, however, treat the system as a black box, as a full technical audit of these architectures is above my paygrade

designed to fulfill a very narrow purpose. It is not a domain general, flexible system. Yet, even then, a complex domain of information has to be involved in the processing taken holistically (including training), for the system of object recognition as the first step in the pipeline.

Furthermore, within processing there is a clear division between three steps of increasing sophistication, but indeed the types of processing we tried to pry apart and delineate. Object recognition, multiframe combination (a type of stacking), and inference-based post-processing to fill in the details not directly available in the input are separate informational operations. It is the latter that endows computational photography with its generative status, according to the engineers responsible for building these systems.

Does it render digital photography partly or mostly generative now? Not necessarily, for as we mentioned, Moon Mode or Scene Optimizer were limited, targeted systems, and they are not exactly commonplace yet (hence the marketing spin on their novelty). But these types of systems proliferate. Google has recently introduced the 'Best Take' (Rakowski et al.,2019[80]) feature in their products, which enables users to replace the unwanted facial expressions of people in the frame, using both the stored tokens of the given person's face and algorithmic post-processing to achieve seamless marquetry. One takes an almost perfect group snap; everyone but a single person is smiling and looking great - the software gives the option to inlay a happy face onto the grimacing dissenter.

These are but gimmicks, limited tricks, a critic could say. That is warranted; these particular systems are concerned with just upgrading on the minor details; one

[80] As with Samsung's case - I am fully aware that this is a company issued blog post announcing the product, not its technical report. But, again, a dive this deep is not necessary for our argumentative purposes. However, to be accommodating to more curious readers, the scientific ancestor of the seamless marquetry employed by the *Best Take* is found in Agarwala et al. (2004).s idea of 'Digital Photomontage'.

could be tempted to classify them as post-processing. But it's not difficult to imagine toy examples where the entire image or at least a major part of the picture is caused by these types of generative engines. Recent successes of diffusion-based architectures (Ho, Jain, Abbeel, 2020) and later transformers, too (Peebles & Xie, 2023) proved it quite forcefully. Creating the entire picture using the text-to-image tools is easy and is ubiquitous. Technology is surely already there. What about philosophy, then? Let's play catch-up and generate (sic!) more cases to probe our intuitions one more time. Then we can finally attempt definitional remedies.

To probe the intuitions and search for differentiating factors for ascribing true generativity, let's consider a thought experiment.

[Albanian Citizen] We present the computational pipeline (e.g., a neural network) a series of tasks that are prompted differently in each case (A-E). In all these five cases, however, regardless of how the instruction is formulated, the system is asked to produce a picture that best realizes the query. Now, let's assume that the system outputs the very same picture in all these cases (100% pixel values equivalence). Say that our end product looks like that:

Figure 17: The final generated output for all cases of [Albanian Citizen]: A-E

Now, let's assume that for all five cases, a different query and/or computational process is used for arriving at the output above.

- Case A - imagine we use a simple a search engine and query 'Albanian citizen'
    - The search engine that looks up the database and brings back the relevant target from the search space (e.g., the internet) that suits the above description best.
- Case B - imagine now, instead of a traditional search engine, we instead inquire a large language general-purpose model and prompt it to generate an 'Albanian citizen'
    - The model then creates the picture from scratch in the sense that it does not load any stored pictorial representation of this picture. It does indeed, plausibly use some sort of internal representation (again - in the deflated, technical sense), but much of a more general nature; it is the effect of relevant learning on general terms such as 'Albania',

‘Citizenship’ or person (but we just assume, the entire process is largely blackboxed).

    - Thus, the result is partly spurious, in the sense that we have not given a direct instruction to generate precisely Dua Lipa in such a setting. Yet, the result is not fully spurious or accidental, because of the more general representation learned through a significant influence of arguably the most famous and online-present living Albanian right now in the training set. Therefore, although the process is optimized for more generalized tasks, the information about Dua Lipa gained during training or learning of the model would still count as boosting performance (increasing the likelihood of generating the picture of Dua Lipa, not some other Albanian - real or hallucinated).

- Case C - now we consider a specialized network trained on images of Dua Lipa, to generate new fictitious tokens of the Albanian in various situations and prompt it to ‘imagine Dua Lipa as getting citizenship.
    - In this case, the model uses the trained representation of an actual person's photos to come up with an ‘unseen’ (not yet existing) picture of that person in a new scenario according to the input instruction.

Now let’s assume that parts of the input are pictorial as well.

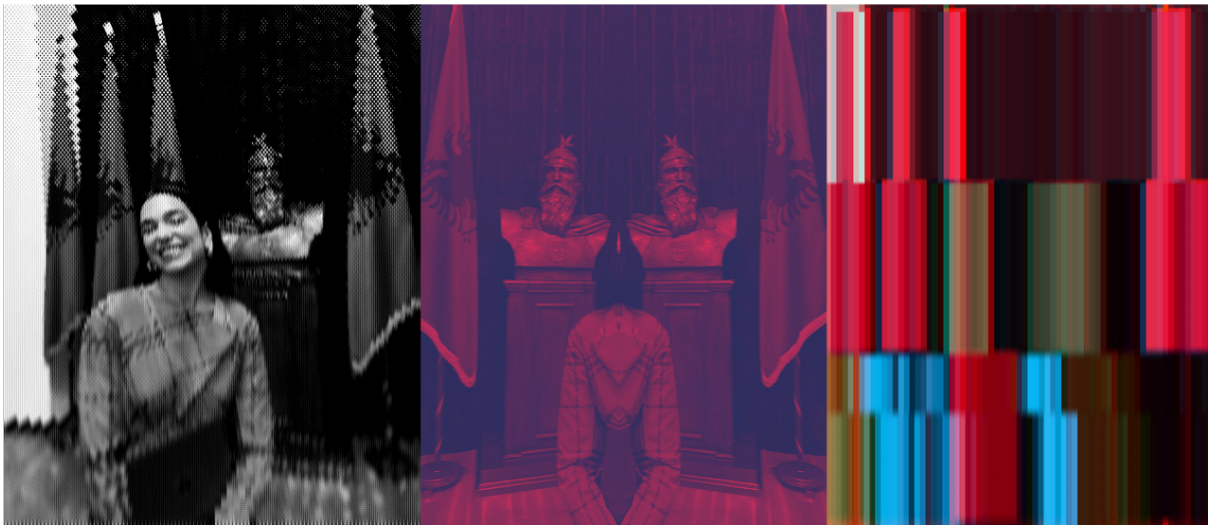

Figure 18: Cases - D-light (limited distortion); D-hard (more lossy distortion with more features lost); E (almost a seed), respectively, left to right.

- Cases D (light and hard - left and center, respectively) are cases of trying to get the model to attempt reversing deformation of the original picture. Imagine we applied a sequence of distorting transforms that have irrevocably deformed the picture, giving us the pictures seen above on the left and center (they differ in this example only by the level of distortion; in the center case, more original information we had lost). Now, using the distorted version, we want to go back to the original form. However, unfortunately, we do not store the history of performed transformations. Neither can we reverse the procedures, because the information was lost and is not expandable from the current format. We try to salvage what we can and, as such, revert to a clever algorithm (a neural network) and feed it the distorted picture as a gist and ask it to make its best informed guess about the original source. It exceeds expectations and achieves full accuracy. It is relatively unlikely, granted, but still not purely accidental - the gist plus known statistical dependencies for natural scenes give the well-trained model enough of a head start to be able to reconstruct the original arbitrarily closely (in principle).

- Case E - finally, let's imagine a case similar to D. Again, we use the 'blob' (right) as a seed for the model to decompress from its latent model space. However, with a crucial difference, we assume here that the blob is not causally connected to the original. Note here that, in the example, I did produce the blob through heavy distortion, so there is still a causal link to the original. But for the sake of the illustration, let's imagine it is a more or less random blob. Now what we do is ask the model to try to predict (guess?) from this non-causal seed; we ask: 'how would this blob look if it was Dua Lipa' and it returns the same output as in all previous cases.

So, in which sense are all these cases different, then? Well, for A, the answer is straightforward. It's a non-generative case, but the output is not recreated, just retrieved. Information is transmitted from the source (the picture's original address) and not created or recreated. But for B-E, the interpretation is more difficult. All of them are tentatively generative, because not all relevant information is present and accessible to the system before it processes the query. As it is known, the loss of information is not reversible solely based on input. Neither is the case of 'unfurling,' where quantitatively more information is produced from a shorter formula according to some rigid rules (as in proc-gen).

But the outcome is informationally richer as well for our tentatively generative cases. Hence, the foundational question for this project is, where does this missing information come from? Obviously, there are two main candidates. The first option - the lacking information comes from the prompt. Surely, the prompt is responsible for the nature of the transformation to be performed by the system. After all, prompts are instructions that should semantically contribute to what type of outcome the operator can expect at the end of the pipeline. But first, as we remember from VAE's example, decompressing from latent space can be triggered by a random seed, bringing about a token different from any original input,

pertaining to the same kind, but not directly instructed by additional information. This suggests that the import of the information from the prompt is not requisite for at least some versions of generative processing. But even for prompted cases, and these are all remaining cases in our example (B-E), the prompt is unlikely to account for the entire informational completion. Information storage from within the model has to contribute, too.

Cases B and C intuitively differ because of the scope of the model, breadth of information that the model has to have internalized and be able to generalize from. Also, the flexibility of them differs - the first one is a general-purpose LLM, whereas the second is a domain-specific, fine-tuned type of model. Plausibly, the type of inference performed, differs, too. Recall the discussion of the importance of locality/globality as a criterion for true extrapolative generalizability. Here we see two systems of significantly different scope, but both relying on informational completion based on the highly-compressed internal model. Both produce new pictures adequately in accord with instructions, regardless of their varied abilities and purposes (the B model is plausibly much more powerful and flexible). This suggests that locality/globality might not constitute a defining feature of generative models.

Case D-light may spring associations with the aforementioned cases of denoising. The seed is recognizably close to the original (and in this case, the target). Still, not enough information has been salvaged for a non-completionist, traditional denoising. It becomes even clearer for the more extreme D-case [D-hard]. There, the level of deformation makes the original content barely recognizable (if at all). The task of the model might be more accurately described as ‘reconstruction’ here than mere denoising. Still, because of the causal link between the original picture and its distortion, the informational operation amounts to the necessity of

performing the recreation of the original signal. The lesson here is simple: regardless of ideal pixel-value identity, the etiology matters.

The importance of the causal structure connecting the input trace (i.e. etiology) to the target becomes more apparent in comparison to case E. As we remember, the blob there is purely uncorrelated with the target picture. There is no reversal to speak of. Although the blob certainly enters the informational picture as a source of information ('inspiration' to cook up such-and-such Albanian citizen that somewhat resembles it), it surely does not suffice to produce the output. As there is no causal lineage to follow, nor is there sufficient prompt input, the responsibility lies on the internal model decompression and its learned regularities that it employs. Case E is another strong candidate for genuine generativity.

But are these cases on a spectrum? Or is it just a family of distinct processes? When is it merely transformed, and when does generativity kick in?

6. Towards the definition

Based on the analysis conducted so far, we are better positioned to tackle the ontology of generativity and try proposing a definition that would better serve the explanatory criteria for genuine novelty that emerged from our analysis of the case studies while conserving some intuitions about genuine generativity (as compared to faux-generativity). By that, one also attempts to demonstrate that only certain architectures achieve generativity (or should be ascribed that ability at least) and that while there are differences between generative pipelines, they are not necessarily constitutive for generativity itself (e.g., non-locality).

What was shown is that the brute generation view is not sufficient, nor are the mathematical methods for enforcing a generation/non-generation boundary, and a

content-based approach cannot be avoided. For the end product to be truly new (generative), it must not be merely outside the learned space (within the space of the already known, informally), but only when there is a content net gain.

Thus, what was crucial across all candidates for genuine generativity was that all pipelines capable of generating genuinely acted upon actively stored information to systematically transform the signal in a regimented way according to these internally represented regularities. This stored information must contentfully correspond to some regularities that enable that filling-in, which provides the missing informational import to infer from. We come back here to the idea we started with inference as informational fill-in whenever there is not enough information present 'online' or 'directly' (the same principle that guides gist+completion paradigms or fuzzy trace theories in memory research; Reyna & Brainerd, 1992; 2002; 2004). In short, previously stored information is mobilized for generation of a new batch of information.

We can sum it up using a slightly worn out paraphrase: no generation without compressed information.

New information tokens are genuinely generated if and only if previously stored information about underlying regularities that are explicitly represented for and accessed by the system are used in virtue of their content.

Note here that this condition might resonate with the discussion on the nature of inference itself, where in Boghossian's groundbreaking paper (2014), one can find the 'Taking Condition' [for an ensuing discussion, see e.g., Wright (2016) or Broome (2013); a major contrary view can be found in (McHugh & Way, 2016)]. [Taking] in its original form posits that for an inference to occur, it is not only required that the premises support the conclusion, but also that the conclusion is drawn because of that taking. In other words, in the virtue of their content and not only in accord

with it. Boghossian originally invokes [Taking] to highlight the role of the agent performing the inference and their intentional usage of information entailed in the premises. For our purposes, the personal-level description and intentional terminology are overmentalistic (and, as promised in the beginning of this chapter, a reductionist, naturalistic account of generativity is argued for here), so the analogy is only partial. Regardless, it conveys an important intuition about the causally active role of informational content of premises to make an inferential move. Another useful intuition pump could be the requirement of being causally active in the virtue of their content, imposed on representations by Ramsey (2007).

Let's bundle up all the restrictions that were drawn for generativity and repack them into a definition.

[Generativity]: A genuinely generative process is a process that fulfills the following set of conditions:

T1: Produces a quantitative net gain batch of information [brute generation]

T2: This new batch of information is non-identical to parts of the training regime (used in the learning of the internal model), nor is it entailed in already known data. [variationality]

T3: But the new output is content-preserving (e.g., subsumed under the same type) [contentfulness]

T4: Production of the output tokens takes advantage of regularities that are actively stored within the compressed internal model, which retains (mirrors) information about the underlying structure of the environment (through learning data) [internal structure]

T5: Generation is performed in virtue of this stored informational content [taking]

T6: Internal informational import is used to **fill in** the information lacking in the direct input (prompt, seed, stimulus) that is available online to the agent [inferential completion]

I argue that the view defended here avoids being too wide and trivializing the notion of generativity, as the naive brute generativity view would. Simultaneously, it avoids being overly restrictive, particularly with being agnostic about personal-level processing necessary for genuine generativity (creativity, novelty), sidestepping both the pitfalls of the traditional philosophical insight of generativity as a phenomenological phenomenon, which declares a priori which agents are capable of performing generative processing, rather than assessing it fairly ex post, based on empirical evidence. For, eventually, if we want to be serious about our naturalistic commitments, we as theorists should strive to provide fair definitional standards that all agents should be judged against to responsibly delineate which types of systems can achieve which types of cognitive processing.

Another upshot of this approach is that it provides a way to potentially escape the type of criticism in the mold of 'stochastic parrots' (Bender et al., 2021) or 'blurry JPGs' (Chiang, 2023) arguments that are too broad and vague to be considered truly damaging. If we want to truly investigate where the boundary lies and what type of system can clear the bar, we should provide a more detailed description of specifications (functional or organizational) that could satisfy the critics. If any artificial information processing systems are excluded solely based on being artificial information processors doing statistical inference, that equals throwing the baby out with the bathwater (even if I agree with their assessment of some current state of affairs in LLMs, the critique is not even too scathing, but too indiscriminate). What we need as theorists are more granular operationalizations and more careful definitions, which I intend to provide here.

Thus, additionally, the account present demonstrates the importance of philosophical theorizing, showing that purely mathematical/formal solutions are not only encumbered by potentially problematic assumptions but also insufficient for explanatory purchase in principle. Defining genuine generativity satisfactorily and prying it apart from faux-generativity must be consistent with philosophical intuitions about the wide array of potential candidates for generativity, which is what the account presented here hopefully has achieved.

I flatly reject critiques that trying to set these boundaries is hopeless because of the heterogeneity of generative systems (Bender & Koller, 2020). Nor do I agree with statements that posing Socratic questions on the ontic status of computing processes are pseudoproblems (Dijkstra, 1984)

On the pragmatic side, ostensibly it makes a lot of sense to be able to discern information retrieval from generation and to assess on which side recapitulation falls. Establishing this definition provides the basis for distinguishing between loaded, processed, and hallucinated images, videos, and texts (see Millière, 2022, on vector grounding). Legal and practical ramifications are not very hard to see and are even more so with the flurry of recent lawsuits on behalf of artists (Andersen v, Stability AI, 2023), journalists (The New York Times Co. v. Microsoft, 2023), academics (Basbanes v. Microsoft, 2024) or even Reddit users (J.L. v. Alphabet, 2023).

On the explanatory side, it seems to me that cognitive science cannot do without the ability to disambiguate between generative mental processes, such as imagining, planning, or counterfactual thinking, and non-generative ones like filtering, synthesizing, and direct perception. Also, we want to be able to assess generativity within debate about particular mental natural kinds. Memory research is a prime example here. Being able to differentiate sharply between conditions for generative and non/faux-generative seems to me as crucial for judging between

constructive, active, completionist (Michaelian, 2011; 2024, De Brigard, 2014) theories and passive, transmissive, retrieval-based theories (see: Martin & Deutscher, 1966 for a classical exposition of *preservationism*; Bernecker (2010) is a more modern defender; Cheng & Werning, 2016 provide a middle-ground proposal) without degenerating into verbal disputes (Chalmers, 2011). The same goes for debates about inferentialism in perception or predictive coding or generative models of cognition in general. One can disagree whether a purely information-based picture suffices to set up and police these boundaries, and that is fair. However, even if that is the case, it is impossible to go further with this explanatory segmentation (e.g., onto the questions on veridicality, causal structure, and functional analysis) without assessing how much can be done on an informational level first. Also, the information-based analysis is inescapable, inasmuch as virtually all approaches in (philosophy of) cognitive (neuro)science subscribe to the idea of information processing as at least a necessary, minimal condition for cognitive processing.

7. Potential Objections

The account of genuine generativity proposed here—defined by inferential completion via compressed internal models—faces challenges from two fundamentally different directions. From "below," critics question whether the computational mechanisms proposed can actually deliver the claimed generative capacity. They ask whether the mathematical and empirical realities of modern AI align with compression principles or if raw scale has rendered efficiency obsolete. From "above," critics challenge the philosophical commitments embedded in the definition itself. They argue that the normative language of "taking" or "following" rules over-intellectualizes what should be a naturalistic account, potentially

excluding the very systems (animals, neural networks) the theory aims to explain. I address these challenges in turn, demonstrating that both sets of objections ultimately clarify rather than refute the compression thesis.

7.1 The paradox of scale: nominal vs. effective complexity

The most pressing empirical challenge arises from the current state-of-the-art in generative artificial intelligence. The compressive generativity commitment - understood that not only optimal learning (e.g., generalization) but also the 'creative' dimension (e.g., generation) - is fundamentally reliant on efficient compressing. This line of thought goes back at least to the Minimum Description Length principle—the idea that learning equates to finding the shortest program capable of reproducing the data (Rissanen, 1978; later also Vitanyi & Li, 2000; recently revised by Grunwald, 2019). This principle suggests that generalization is a function of compression: the more compact the model, the better it captures the underlying regularities.

However, the large language models (LLMs) that display the most impressive generative capabilities are defined by massive over-parameterization: they possess far more trainable parameters than data points in their training sets. Classical statistical learning theory predicts that such excess capacity should cause catastrophic overfitting—the memorization of noise rather than the extraction of structure. Yet, empirically, we observe the opposite.

The phenomenon of double descent (Belkin et al.,2019) demonstrates a counterintuitive trajectory. While test error initially increases as models grow larger (the classical overfitting regime), it eventually decreases again once models become sufficiently over-parameterized. Belkin and colleagues demonstrate this pattern across diverse architectures, from decision trees to deep ResNets. If the most generative systems known to science are massive and redundant, the claim

that "generativity requires compression" appears empirically falsified at precisely the frontier where it should matter most.

Furthermore, the "grokking" phenomenon (Power et al.,2022) suggests that generalization often emerges long after the training set has been ostensibly memorized. In experiments with modular arithmetic, networks achieve zero training loss quickly—memorizing the dataset—but remain at near-random test accuracy. However, if training continues for thousands of additional epochs, the model suddenly "groks," transitioning to perfect generalization. This suggests that generativity emerges not from efficient coding during the learning phase, but from extended exploration of the solution space. If sheer scale and extended optimization drive the transition from memorization to generation, the Neo-IPP loses its explanatory target.

Response:

This objection is powerful, but it rests on a confusion between nominal complexity (the hardware size or parameter count) and effective complexity (the structure of the solution actually implemented). Recent theoretical work vindicates the compression view by revealing that over-parameterization is not an alternative to compression, but the most efficient means to discover it.

The core idea lies in the geometry of the loss landscape. As shown by Hochreiter and Schmidhuber (1997) and rigorously confirmed by Mingard et al. (2021), neural networks do not settle into just any zero-error solution. Instead, stochastic gradient descent (SGD) is biased toward "flat minima"—large connected regions where the error function remains constant under parameter perturbation.

Crucially, these flat minima correspond to solutions with low Kolmogorov complexity. A solution that is robust to parameter perturbation requires fewer bits

to specify than a sharp, isolated solution, because it does not demand high-precision weights. Thus, while the network possesses billions of parameters, the functional circuit it implements at convergence is a simple, highly compressed program. The massive size of the network is not the solution itself; it is the search space that enables the optimization process to locate this compressed circuit. Over-parameterization provides the high-dimensional "room" required for SGD to navigate around local minima and find the low-complexity attractor.

This interpretation is directly confirmed by mechanistic interpretability research into grokking. Liu et al. (2022) demonstrate that the "grokking" moment coincides with the network transitioning from a high-complexity "memorization" circuit to a low-complexity "algorithmic" circuit. Nanda and Lieberum (2023) bolster this even more: they show that post-grok networks implement compact algorithms rather than lookup tables. The network does not generate because it is large; it generates because its scale allows it to find the compressed structure implicit in the data. Thus, the empirical reality of deep learning supports the Neo-IPP: generation is always the product of a compressed representation, even if the machinery required to find that representation is vast.

7.2 Biological implausibility

A second, perhaps more intuitive objection targets the biological plausibility of a "compression-only" cognitive architecture. While semantic memory and conceptual knowledge are readily understood as compressed schemas, the biological brain appears to violate the compression mandate in other faculties, such as episodic memory.

Human episodic memory is phenomenologically characterized by the one-shot retention of highly specific, high-dimensional details—the exact hue of a birthday cake, the precise timbre of a parent's voice, or the spatial texture of a specific

afternoon. These memories often feel like "raw storage": high-fidelity recordings of sensory experience that can be replayed with vividness and specificity. If the brain relies on such raw storage for its simulation capacities, then the "compression-only" account of generativity is biologically insufficient. It would suggest that evolution maintains a metabolically expensive dual-track system: one compressed (for concepts) and one raw (for episodes). If genuine generativity requires compression, why does our most sophisticated memory system seem to prioritize fidelity over abstraction?

Response:

The "raw storage" view of episodic memory, however compelling it may feel to the subject, is a folk-psychological illusion systematically undermined by empirical research. Far from providing a counterexample, episodic memory exemplifies the core principles of generative reconstruction from compressed gist. The vividness of recall is not evidence of the fidelity of storage, but of the power of the generative model doing the reconstruction.

The Deese-Roediger-McDermott (DRM) paradigm provides the canonical demonstration (Roediger & McDermott, 1995). When participants study semantically related word lists (e.g., nurse, hospital, medicine, patient), they routinely "remember" the non-presented critical lure ("doctor") with high confidence. Crucially, they do not merely report the lure as familiar; they generate rich phenomenological details of its occurrence—recalling its specific position in the list or the voice that spoke it (Pardilla-Delgado & Payne, 2017). This robust illusion proves that the system does not store a verbatim record of the event. Instead, it stores a compressed gist (the semantic theme) and generates the details upon retrieval. The vividness of the "memory" tracks the coherence of the simulation, not the accuracy of the trace.

This aligns with the Constructive Episodic Simulation hypothesis (Schacter & Addis, 2007). Neuroimaging studies demonstrate that remembering the past and imagining the future recruit the same core neural network (hippocampus, medial prefrontal cortex). This shared substrate implies a shared computational mechanism: both processes reconstruct specific episodes by sampling from learned distributions and schemas. We do not store "video files" of experience; rather, we store semantic pointers (Eliasmith 2013,)—compact, low-dimensional codes that serve as indices (as argued in Ch. 6).

The Proustian madeleine is not a file handle for a stored recording; it is a compressed seed. When triggered, this seed initiates a pattern completion process in hippocampal attractor networks (Norman & O'Reilly, 2003). The partial cue reactivates a subset of neurons, and the network's recurrent dynamics—shaped by the compressed statistical regularities of past experience—"fill in" the missing information to regenerate a stable, high-dimensional state. The Fuzzy Trace Theory (Brainerd & Reyna, 2002) further supports this by showing that memory judgments are driven more by "gist traces" than "verbatim traces" as time passes. Thus, episodic memory is not the retrieval of raw data but the generative decompression of a sparse trace using the brain's internal models.

7.3 A view from the top: philosophical objections

Finally, the account faces serious philosophical challenges from "above." These objections target the normative language used to define generativity—specifically the requirement (derived from Boghossian's Taking Condition) that genuine generation occurs "in virtue of the content" of internal representations. This commitment appears to over-intellectualize the process, potentially excluding the very systems the theory aims to explain.

The “over-intellectualization” objection argues that if generativity requires an agent to consciously "take" a rule or "follow" a norm, then paradigm cases of generative behavior are excluded. Consider a three-year-old who generates the novel sentence "I goed to the park." The child is clearly applying a generative rule (add -ed for past tense) to produce a novel output, but they do not "take" this rule to support the conclusion in any recognizable, personal-level sense. Similarly, a brainy crow that spontaneously combines sticks to solve a puzzle is generating a novel solution, but it lacks the metacognitive machinery to represent inferential support relations. If the Neo-IPP requires "taking," it seems to restrict generativity to adult human reasoning, leaving the vast majority of biological novelty unexplained.

Closely related is the counterargument that we can call “the level-crossing objection.” Critics might argue that applying terms like "inference" or "generation" to subpersonal neural circuits commits a category error. As articulated by Drayson (2014), the distinction between the personal level (agents, reasons) and the subpersonal level (mechanisms, causes) is fundamental. If generativity is defined by subpersonal mechanisms (compression algorithms), calling it "reasoning" is mere anthropomorphism. Conversely, if it requires personal-level agency, then neural networks cannot be generative. The account seems trapped in a dilemma: either it trivializes generativity by reducing it to mere causal processing, or it inflates it by demanding impossible metacognition from simple systems.

Response:

These objections rest on a conflation of the personal and subpersonal levels of explanation. The Neo-IPP adopts a functional-computational reading of the "in virtue of" relation, not a personal-level metacognitive one.

As argued by Ramsey (2007), for a state to function as a representation, it must play a causal role in virtue of its content. At the subpersonal level, this does not require

the agent to be aware of the content; it requires the system's computational transitions to be structurally isomorphic to the semantic rule. When Olah et al. (2020) identify "induction heads" in large language models, they identify a physical circuit that functionally implements a copying rule. The system does not "know" the rule, but the causal structure of the hardware ensures that the output is counterfactually dependent on the content of the representation. If the content were different, the output would be different.

This functional reading resolves the dilemma by treating generativity as a capacity-based rather than a level-based phenomenon (Frankish, 2009; Shea, 2021). Generativity is defined by the architectural capacity to compress experience into portable schemas and decompress them into novel simulations (T1–T6). This computational architecture can be embedded at the subpersonal level—as in the automatic pattern completion of the crow's sensorimotor cortex or the child's morphological processing—where it operates without awareness. It can also be embedded at the personal level, where it is integrated with metacognitive access to enable conscious, deliberate reasoning.

The difference isn't whether generation occurs, but how it is embedded in the cognitive architecture. The crow generates novel tool combinations subpersonally; the engineer generates novel designs personally. Both rely on the same underlying compression-decompression architecture, but only the engineer has metacognitive oversight. Thus, the "in virtue of content" clause is not a demand for hyper-intellectualism; it is the naturalistic criterion that distinguishes genuine computation from mere physical reflex.

These objections—targeting computational scale, biological memory, and philosophical agency—ultimately strengthen the compression thesis. The paradox of scale reveals that compression is the destination of learning, even if

over-parameterization is the necessary vehicle. The biological challenge confirms that even the most vivid memories are generative reconstructions, not raw playbacks. And the philosophical challenge paves the way for clarifying that generativity is a fundamental computational capacity that scales from the silent circuits of vision to the deliberate reasoning of the scientist. Far from refuting the Neo-IPP, these challenges demonstrate that compression is the unified principle that makes the generation of information possible in a physical world.

8. Conclusions

In this chapter, I argued for a novel account of generative processing. While the problem has largely (and surprisingly) avoided more detailed attention from philosophers. To ameliorate that, I have suggested a new account of generative processing based on the idea of internal inferential completion (aligning with constructive memory accounts: Michaelian, 2016; De Brigard, 2014; Robins, 2020). It draws heavily from both philosophical intuitions and recent empirical findings from both cognitive science and artificial modelling in neural networks research. I assessed some potential criticisms and provided tentative responses to them. The chapter is concluded with a short defense of both the reality and importance of defining generativity both for pragmatic purposes and for the central (and so far, unfortunately, mostly tacit) role it plays as an explanatory concept in cognitive science.

Final Conclusions

1. Philosophy In and Of Cognitive (Neuro)science

This dissertation has operated on two parallel tracks. On the one hand, it is a work in the philosophy of cognitive science—a meta-analytic project designed to audit the explanatory strategies of the field, defend the Information Processing Paradigm (IPP) against its modern critics, and regiment the notoriously "messy" and heterogeneous literature on information. By distinguishing between f-information and a-information and by cataloging the operational logic of neural computation, we have attempted to impose conceptual order on a domain often plagued by metaphorical looseness.

On the other hand, this project is an exercise in philosophy in cognitive (neuro)science. We have not been content to merely spectate from the armchair; rather, we have aimed to contribute substantive theoretical tools—such as the definition of "genuine generativity" or the "Correspondence Network Framework"—that can be deployed by working scientists to solve specific empirical puzzles. The defense of the Neo-IPP is therefore not an act of nostalgia for the computationalism of the 1970s, but a forward-looking proposal for a "responsible mindreading" program that integrates the rigors of philosophy with the operational realities of neuroscience.

Finally, a brief note on the methodological stance of this dissertation. Alex Morgan (2020) cautions against the perils of 'armchair neuroscience'—the tendency of philosophers to dictate from the comfort of their studies how the brain *must* function based solely on computational logic, much like an *internet quarterback* shouting plays at a professional athlete who is actually on the field, getting hit. While this project is, by necessity, a product of the armchair—my primary

laboratory equipment being a laptop and an ungodly amount of procrastination—it aims to model a 'responsible armchair' approach. Rather than deducing biology from first principles, I have sought to synthesize the logic that biology has already revealed. By grounding our definitions of format and content in the empirical details of dendritic computation (Phillips) and representational geometry (RSA) and artificial modelling (generative neural nets), the goal is to ensure that philosophical analysis serves not as a substitute for empirical data but as a framework for organizing it. The armchair remains, but hopefully, the shouting is a bit more informed.

To check that, let's revisit the empirical meat on this conceptual skeleton.

2. The Stress Tests Recap

In Part 2, we subjected the core commitments of the Neo-IPP to three empirical "stress tests." These were not chosen at random, nor do they function as independent case studies. Rather, they map the sequential trajectory of an intelligent system: how it structures the world (compression), how it validates that structure (matching), and how it exploits that structure to go beyond the given (generativity). I revisit briefly these parts, which were not summarized yet (chapters 1-5 were summed up after Part 1)

Representing as Compressing (Chapter 6)

We began by arguing that the format of representation is not arbitrary. Driven by severe metabolic constraints and the "curse of dimensionality," cognitive systems are forced to compress the environmental signal. However, Chapter 6 demonstrated that this compression is not merely a data-saving measure; it is an engine of ontology. By maximizing the scaffolding-to-output ratio, the brain does not just shrink the signal; it disentangles the causal structure of the world. Efficient

coding, therefore, is the mechanism by which the "blooming, buzzing confusion" (1890:488) of sensory input is transmuted into the tractable formats of neural representation.

Meaning from Matching (Chapter 7)

If compression explains the origin of the vehicle, structural correspondence explains the fixation of content. In Chapter 7, we relied on Representational Similarity Analysis (RSA) to bridge the gap between the scientist's measure and the brain's operation. We argued that "representational geometries" are not observer-relative artifacts ("blobology"), but real, causally efficacious structures that are preserved across different physical substrates. The correspondence network framework reveals that content is not a mysterious "gloss" added to the system, but a functional property of systems that match the relational structure of their internal states to the relational structure of the world.

Genuine Generativity (Chapter 8)

Finally, in Chapter 8, we addressed the most pressing and least understood capacity of the cognitive system: generativity. While the current zeitgeist is dominated by "generative AI," the theoretical literature has lacked a rigorous criterion for distinguishing true generation from mere retrieval or procedural mimicry. This task can be understood as laying naturalistic conditions for creativity from an informational standpoint. We filled this gap by defining the "interpolation/extrapolation boundary." We argued that genuine generativity—the ability to produce novel, plausible information that lies outside the convex hull of the training data—is the ultimate functional test of the Neo-IPP. It requires more than statistical correlation; it requires the system to actively store and manipulate compressed, contentful schemas. In doing so, we provided a criterion that allows us

to distinguish the creative extrapolation of a planning agent from the "stochastic parroting" of a mere interpolator.

3. The Explanatory Cascade Revisited

Viewed together, these chapters do not merely defend the paradigm; they describe a unified processing pipeline. We can conceptualize this as an explanatory cascade (which a sample of we encountered first in Chapter 5, section 3), where each stage provides the necessary conditions for the next:

1. Causal grounding: the system is causally coupled to the environment.
2. Structural format: metabolic pressures force the system to compress this causal input into efficient, disentangled formats (source coding).
3. Semantic content: these formats are validated through structural matching (relevant infocorrespondence), fixing their content.
4. Generative behavior: finally, these content-bearing, structured vehicles are manipulated to generate novel behaviors and simulations (extrapolation).

The Neo-IPP, therefore, is not a static list of commitments but a description of this cascade. It explains how a physical system moves from the raw causal impact of the world to the creative generation of new ideas.

4. The Engine of Truth: Verisimilitude and the Consumer

If the explanatory cascade describes how the system operates, we must still answer why it converges on these specific solutions. We argue that the ultimate driver of this convergence is not merely metabolic efficiency, but verisimilitude. The brain is an engine of truth approximation. As rate-distortion theory insists, the "sweet spots" of efficient coding are not random artifacts; they align with the natural kinds (homeostatic property clusters) of the environment. Formats like polar coordinates

or hierarchical vectors are selected because they "carve nature at its joints," minimizing the distortion between the internal model and the external reality. Veridicality pays off because it is the only strategy that minimizes prediction error in the long run. Thus, the drive for compression, the matching of geometries, and the capacity for portability are all manifestations of a singular biological imperative: to construct a map that is structurally faithful to the territory.

4,1 Evading the Empty Symbols (via Consumer Mechanisms)

This focus on verisimilitude allows us to finally dissolve the empty symbol problem that has plagued computationalism since Searle (1980). Critics like Bickhard and Terveen (1995) as well as Brette (2019) have rightly argued that a symbol without a consumer is causally inert—a mere gloss by the observer. We agree. However, the Neo-IPP does not rely on floating vectors. We solve the empty symbol problem by positing the necessary existence of a consumer system.

A symbol is not "empty" if there exists a downstream mechanism evolved or learned to exploit its specific format. As the "decoder's dilemma" reminds us, information is only f-information (information-for-the-system) if it is formatted in a way that makes it available (Kulvicki) to this consumer. The algorithmic homunculus is not a ghost in the machine but a physically instantiated circuit (like the induction heads in LLMs or the decision boundaries in PFC) that "reads" the compressed vector and transforms it into action. By grounding content in the specific requirements of these consumers, we close the causal loop without magic.

4.2 The Responsible Mindreader: Aligning f-info and a-info

This distinction has profound implications for the practice of neuroscience. For decades, the field has struggled with the gap between the scientist's statistical measures (a-information) and the brain's actual operations (f-information). The

Neo-IPP provides the theoretical bridge. Responsible mindreading—the project of decoding mental content from neural activity—is not an exercise in phrenology or blobology. It is the rigorous attempt to align our external decoding algorithms (like RSA or MVPA) with the internal decoding mechanisms of the brain. When we successfully decode a mental state, we are not imposing a metaphor; we are discovering the specific structural format that the brain's own consumers are using to drive behavior.

4.3 The forgotten dimension: operations

Finally, our analysis highlights a critical blind spot in standard philosophical accounts: the neglect of neural operations on information. Philosophers often focus on the vehicles of representation (vectors, symbols) while ignoring the operations that manipulate them. Drawing on the work of Phillips and others, we have emphasized that mechanisms like apical amplification, divisive normalization, and coincidence detection are not just implementation details; they are the functional primitives of cognition. Philosophy of cognitive (neuro)science must engage deeply with this operational logic if it hopes to explain how abstract computations are realized in wetware.

5. The explanatory landscape: pluralism, not war

To stick to the figure of speech employed throughout this dissertation, where does this leave the "Siege" on the paradigm? We argue that the supposed war between computationalism and its rivals is largely a failure of partition. The Neo-IPP advocates for methodological pluralism. We do not seek to eliminate dynamical systems theory (DST) or causal-mechanistic explanations; rather, we cooperate with them. As demonstrated in our analysis of the two-point processor, mechanistic details (biophysics) are not the enemy of computation; they are its

implementation. Similarly, dynamical descriptions are indispensable for explaining the continuous evolution of the system's state space. However, we insist that these strategies are insufficient on their own to explain high-level cognition. Without the postulate of content, dynamics is blind; without the postulate of format, mechanism is mute.

This cooperative stance, however, has a limit. Against the radical claim of enactivism—that basic minds are contentless and representation-free—we stand firm. The enactivist reliance on dynamic coupling is sufficient to explain the online, reactive coping of a Watt governor or a phototactic moth. But it crumbles when faced with the problem of decoupling. As we argued in Chapter 2 and demonstrated in Chapter 8, the capacity to solve representation-hungry problems—to plan for the future, to reason about the absent, and to generate novel solutions (roughly: extrapolation)—requires the system to break the causal chain of the here-and-now. It requires a vehicle that can stand in for the world. To deny this is to render the most distinctive features of cognition inexplicable.

6. The Fodorian Phoenix

We conclude, therefore, that the reports of the death of the information processing paradigm have been greatly exaggerated. The paradigm has not died; it has matured. It has evolved from the brittle computer metaphor of the 1970s into a robust, biologically plausible framework for understanding how physical systems become epistemic engines.

In the early days of the cognitive revolution, Jerry Fodor famously championed the information flow psychologists against the behaviorist orthodoxy, arguing that we could not explain the intelligence of behavior without positing internal vehicles that carry, transform, and preserve information. For a time, under the siege of enactivism and dynamicism, it seemed this view might be swept away by the focus

on continuous coupling and environment-body dynamics. But as we have argued, while dynamics explain the flow, they cannot explain the content.

The Neo-IPP thus represents a Fodorian Phoenix. It rises from the ashes of GOFAI not to declare war on the new mechanistic and dynamical sciences, but to provide them with the semantic soul they otherwise lack. We accept the pluralist insight: the brain is indeed a dynamical system and a biological mechanism. But to explain how that mechanism solves representation-hungry problems—how it tracks truth, filters noise, and generates novelty—we must return to computational insight. As Fodor himself presciently observed at the dawn of the paradigm:

"It is, in short, the claim that mental processes are computational processes... [and] that the study of the mind is the study of the processing of information." (Fodor, 1975).

By grounding information in verisimilitude, formatting it for consumption, and validating it through structural correspondence, the Neo-IPP provides the coherent explanation for how a collection of neurons can capture the causal structure of the universe and use it to generate new worlds. In the face of the siege, we do not retreat.

Crucially, however, the survival of this paradigm does not hinge on the permanence of the specific empirical exemplars mobilized in these pages. While I have leaned heavily on mechanisms like apical amplification and methodologies like RSA to ground the theory, these serve primarily as functional proofs of concept—demonstrations that the explanatory cascade is biologically plausible —a validation of 'semantics without magic' (see: Conclusions of Chapter 4) — not that its current mapping is immutable[81]. The conceptual strength of the paradigm lies in

[81] I have hinted at this in the footnote to Lance Nizami's critique of the shattered landscape of neuroscientific implementations of information theory currently on offer (in section 2.1, ch 1).

the logic of the cascade itself—the necessity of distinguishing format from content and validation from generation—rather than in the specific biological hardware we currently identify as its implementation. Even if future neuroscience were to upend our current models—perhaps eventually vindicating Gallistel's search for digital gates and read/write engrams over the analog mechanisms favored here—the core commitments of the Information Processing Paradigm would endure. This resilience is precisely what identifies the Neo-IPP not as a static dogma, but as an evolving lineage. Just as this framework emerged as a necessary second iteration—an update to the original classical computationalism adapted to the constraints of the neural age—it would be neither unlikely nor catastrophic if it were not the last; the paradigm will survive the evolution of its details.

The information processing stance may no longer be the "only" game in town—it now shares the board with the neomechanists, DST and causal theories—but it remains the *winning* game. One simply cannot deflate or eliminate information-based explanations from the final formula. In the final analysis, for any science that aspires to explain not just how organisms move, but how they know, this insight remains the indispensable core (as per the quote from Dunn that opens the thesis).

We assert that for any science of mind that aspires to explain not just how organisms move, but how they know, the *informational stance*[82] remains—and will

---

[82] This phrasing intentionally echoes Daniel Dennett's (1987) concept of the intentional stance but with a critical distinction. Dennett famously categorized information processing explanations under the "design stance," a level of abstraction concerned with sub-personal engineering and functional blueprints, distinct from the instrumentalist heuristic of the intentional stance. By adopting an informational stance, the Neo-IPP aligns itself with this engineering realism. We are not merely predicting the system *as if* it processes information; we are identifying the design principles (compression, format, correspondence) that physically constitute its operation. Thus, unlike the intentional stance, which can be agnostic about internal vehicles, the informational stance is committed to realism about vehicles and information processing operations.

likely remain—the indispensable way of explaining and understanding how both the mind and the brain work (they're the same thing).

References

Aaronson, S. (2014, May). *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*. https://scottaaronson.blog/?p=1799

Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Blackwell Pub.

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., & Cohen, M. (2004). Interactive digital photomontage. *ACM SIGGRAPH 2004 Papers*, 294–302. https://doi.org/10.1145/1186562.1015718

Alicea, B., Cialfi, D., Lim, A., & Parent, J. (2024). A Primer on Gibsonian Information. *arXiv preprint arXiv:2403.18829*.

Allen-Zhu, N., Z., Gelashvili, R., Micali, S., &. Shavit. (2014). *Sparse sign-consistent Johnson–Lindenstrauss matrices: Compression with neuroscience-based constraints. Proceedings of the National Academy of Sciences, 111 (47), 16872–16876*.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, *33*(4), 245–266. https://doi.org/10.1017/S0140525X10000853

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The “oblique effect” in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Asdal, K., & Jordheim, H. (2018). Texts on the Move: Textuality and Historicity Revisited. *History and Theory*, *57*(1), 56–74. https://doi.org/10.1111/hith.12046

Attneave, F. (1954). *Some informational aspects of visual perception. Psychological Review, 61 (3), 183.*

Attwell, D., & Laughlin, S. B. (2001). An Energy Budget for Signaling in the Grey Matter of the Brain. *Journal of Cerebral Blood Flow & Metabolism*, *21*(10), 1133–1145. https://doi.org/10.1097/00004647-200110000-00001

Artiga, M. (2016). Liberal Representationalism: A Deflationist Defense. *Dialectica*, *70*(3), 407–430. https://doi.org/10.1111/1746-8361.12152

Auksztulewicz, R., Friston, K. J., & Nobre, A. C. (2017). Task relevance modulates the behavioural and neural effects of sensory predictions. *PLoS biology*, *15*(12), e2003143.

Badcock, J., P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., &. Hohwy. (2019). *The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. Cognitive, Affective, & Behavioral Neuroscience, 19 (6), 1319–1351. doi: 10.3758/s13415-019-00721-3.*

Baggio, G. (2018). *Meaning in the Brain. MIT Press.*

Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*. http://www.staroceans.org/stable-diffusion/2110.09485.pdf

Barack, D., & Jaegle, A. (2019). Codes, functions, and causes: A critique of Brette's conceptual analysis of coding. *The Behavioral and Brain Sciences*, *42*, e219. https://doi.org/10.1017/S0140525X19001407

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371. https://doi.org/10.1038/s41583-021-00448-6

Bar-Hillel, Y., & Carnap, R. (1953). Semantic Information. *The British Journal for the Philosophy of Science*, *4*(14), 147–157. https://doi.org/10.1093/bjps/IV.14.147

Barlow, H. B. (1972). *Single units and sensation: A neuron doctrine for perceptual psychology? Perception, 1 (4), 371–394*.

Barnard, E., & Wessels, L. A. (1992). Extrapolation and interpolation in neural network classifiers. *IEEE Control Systems*, *12*(5), 50–53. https://doi.org/10.1109/37.158898

Barwise, J., J., &. Seligman. (1997). *Information flow: The logic of distributed systems. Cambridge; New York: Cambridge University Press*.

Bashivan, J. J., P., Kar, K., &. DiCarlo. (2019). *Neural population control via deep image synthesis. Science, 364 (6439), eaav9436*.

Bates, R. A., C. J., &. Jacobs. (2020). *Efficient data compression in perception and perceptual memory. Psychological Review*.

Bechtel, O., W., &. Shagrir. (2015). *The Non-Redundant Contributions of Marr's Three Levels of Analysis for Explaining information processing Mechanisms. Topics in Cognitive Science, 7 (2), 312–322*.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines*, *4*(1), 1–25. https://doi.org/10.1007/BF00974201

Bechtel, W. (1998). Representations and Cognitive Explanations: Assessing Dynamicist's Challenge in Cognitive Science. *Cognitive Science*, *22*(3), 295–318.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. New York: Routledge*.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A*, *41*(3), 321–333. https://doi.org/10.1016/j.shpsa.2010.07.003

Beck, J. (2018a). Analog mental representation. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(6), e1479. https://doi.org/10.1002/wcs.1479

Beck, J. (2018b). Marking the Perception–Cognition Boundary: The Criterion of Stimulus-Dependence. *Australasian Journal of Philosophy*, *96*(2), 319–334. https://doi.org/10/gc92cf

Bellman, R. (1957). A Markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Ben-Av, M. B., & Sagi, D. (1995). Perceptual grouping by similarity and proximity: Experimental results can be predicted by intensity autocorrelations. *Vision research*, *35*(6), 853–866.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bergstrom, M., C. T., &. Rosvall. (2011). *The transmission sense of information. Biology & Philosophy, 26 (2), 159–176*.

Bernecker, S. (2010). *Memory: A philosophical study*. Oxford University Press.

Bickhard, M. H. (2000). Information and representation in autonomous agents. *Cognitive Systems Research*, *1*(2), 65–75. https://doi.org/10.1016/S1389-0417(99)00007-8

Bickhard, M. H. (2009a). Interactivism: A manifesto. *New Ideas in Psychology*, *27*(1), 85–95. https://doi.org/10.1016/j.newideapsych.2008.05.001

Bickhard, M. H. (2009b). The interactivist model. *Synthese*, *166*(3), 547–591. https://doi.org/10.1007/s11229-008-9375-x

Bickhard, M. H. (2015). What could cognition be if not computation...Or connectionism, or dynamic systems? *Journal of Theoretical and Philosophical Psychology*, *35*(1), 53–66. https://doi.org/10.1037/a0038059

Bickhard, M. H. (2017). Information, Representation, Biology. *Biosemiotics*, *10*(2), 179–193. https://doi.org/10.1007/s12304-017-9296-5

Bickhard, M. H., & Richie, D. M. (1983). *On the nature of representation: A case study of James Gibson's theory of perception*. Praeger.

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. North-Holland.

Bielecka, K. (2025). *What is Misrepresentation? The role of epistemic accuracy in cognition* (K. Bielecka & M. Miłkowski, transl.). Springer Nature.

Bielecka, K. &. Miłkowski M. (2020). *Error Detection and Representational Mechanisms. In J. Smortchkova, K. Dołęga, & T. Schlicht (Eds.), What are Mental Representations? (Pp. 287–313). New York: Oxford University Press*.

Birch, J., & Smortchkova, J. (2019). From the "coding metaphor" to a theory of representation. *Behavioral and Brain Sciences*, *42*, e220. https://doi.org/10.1017/S0140525X19001456

Blinowska, M., K. J., Kuś, R., &. Kamiński. (2004). *Granger causality and information flow in multivariate processes. Physical Review E, 70 (5), 050902*.

Block, N. (2014). *Seeing-as in the light of vision science. Philosophy and Phenomenological Research, 89 (3), 560–572*.

Block, N. (2023). *The border between seeing and thinking*. Oxford University Press. https://library.oapen.org/handle/20.500.12657/92405

Boghossian, P. (2014). *What is inference?*

Bollen, K. A. (1989). A New Incremental Fit Index for General Structural Equation Models. *Sociological Methods & Research*, *17*(3), 303–316. https://doi.org/10.1177/0049124189017003004

Bonnasse-Gahot, L. (2022). *Interpolation, extrapolation, and local generalization in common neural networks*, arXiv. https://doi.org/10.48550/arXiv.2207.08648

Borst, F. E., A., &. Theunissen. (1999). *Information theory and neural coding. Nature Neuroscience, 2 (11), 947–957.*

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., & Heaton, R. F. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Brain-computer interface (BCI) technology, safeguards, cybersecurity, brainjacking, John Donoghue, Eb Fetz,. (b.d.). Downloaded 15 Dec 2025, https://www.neurotechreports.com/pages/publishersletterJun17.html

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, *10*(1), 3–47.

Brainerd, C. J., & Reyna, V. F. (2002a). Fuzzy-Trace Theory and False Memory. *Current Directions in Psychological Science*, *11*(5), 164–169. https://doi.org/10.1111/1467-8721.00192

Brainerd, C. J., & Reyna, V. F. (2002b). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in child development and behavior*, *28*, 41–100.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review*, *24*(4), 396–439.

Brandom, R. (2013). Global anti-representationalism. *Expressivism, pragmatism and representationalism*, 85–111.

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, *42*, E215. https://doi.org/10.1017/S0140525X19000049

Broadbent, D. E. (1958). *Perception and communication* (s. v, 340). Pergamon Press. https://doi.org/10.1037/10037-000

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Broome, J. (2013). *Rationality through reasoning*. John Wiley & Sons.

Buckner, C. (b.d.). Rational Inference: The Lowest Bounds. *Philosophy and Phenomenological Research*, n/a-n/a. https://doi.org/10.1111/phpr.12455

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, *195*(12), 5339–5372. https://doi.org/10.1007/s11229-018-01949-1

Buckner, C. (2021). A forward-looking theory of content. *Ergo an Open Access Journal of Philosophy*, *8*. https://par.nsf.gov/biblio/10451747

Buckner, C. (2022). *A Forward-Looking Theory of Content. Ergo: An Open Access Journal of Philosophy, 8, 37. doi: 10.3998/ergo.2238*.

Burnston, D. C. (2016a). A contextualist approach to functional localization in the brain. *Biology & Philosophy*, *31*(4), 527–550. https://doi.org/10.1007/s10539-016-9526-2

Burnston, D. C. (2016b). Computational neuroscience and localized neural function. *Synthese*, *193*(12), 3741–3762. https://doi.org/10.1007/s11229-016-1099-8

Burnston, D. C. (2021a). Contents, vehicles, and complex data analysis in neuroscience. *Synthese*, *199*(1), 1617–1639. https://doi.org/10.1007/s11229-020-02831-9

Burnston, D. C. (2021b). Getting over Atomism: Functional Decomposition in Complex Neural Systems. *The British Journal for the Philosophy of Science*, *72*(3), 743–772. https://doi.org/10.1093/bjps/axz039

Butterfill, S. A., & Sinigaglia, C. (2014). Intention and motor representation in purposive action. *Philosophy and phenomenological research*, *88*(1), 119-145.

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press.

Cantwell-Smith, B. (1996). *On the origin of objects. MIT Press*.

Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, *85*, 101200.

Carandini, D. J., M., &. Heeger. (2012). *Normalization as a canonical neural computation. Nature Reviews Neuroscience, 13 (1), 51*.

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought*. Clarendon Press ; Oxford University Press.

Chaitin, G. J. (1977). *Algorithmic information theory*.

Chemero, A. (2000). Anti-Representationalism and the Dynamical Stance. *Philosophy of Science*, *67*(4), 625–647. https://doi.org/10.1086/392858

Chemero, A. (2003). Information for perception and information processing. *Minds and Machines*, *13*, 577–588.

Chemero, A. (2009a). *Radical embodied cognition. Cambridge, MA: MIT Press*.

Chemero, A. (2009b). *Radical Embodied Cognitive Science*. The MIT Press.

Chiang, T. (access: 9th Feb 2023). ChatGPT Is a Blurry JPEG of the Web. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153. https://doi.org/10.1007/s11229-013-0369-y

Chirimuuta, M. (2018). Explanation in Computational Neuroscience: Causal and Non-causal. *The British Journal for the Philosophy of Science*, *69*(3), 849–880. https://doi.org/10/gd33hv

Chirimuuta, M. (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(01), 1–1. https://doi.org/10.1017/S0140525X00001515

Christensen, W. (2010). The Decoupled Representation Theory of the Evolution of Cognition—A Critical Assessment. *The British Journal for the Philosophy of Science*, *61*(2), 361–405. https://doi.org/10.1093/bjps/axp043

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1585–1599. https://doi.org/10.1098/rstb.2007.2054

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception & Psychophysics*, *81*(7), 2265–2287. https://doi.org/10.3758/s13414-019-01760-1

Clark, A. (1992). Presence of a Symbol. *Connection Science*, *4*(3–4), 193–205. https://doi.org/10.1080/09540099208946615

Clark, A. (1997). *Being there. MIT Press Cambridge, MA*.

Clark, A. (2013). *Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36 (3), 181–204. doi: 10.1017/S0140525X12000477*.

Clowes, M. B. (1971). On seeing things. *Artificial intelligence*, *2*(1), 79–116.

Cohen, J., & Meskin, A. (2006). An objective counterfactual theory of information. *Australasian Journal of Philosophy*, *84*(3), 333–352. https://doi.org/10.1080/00048400600895821

Cohen, A. R., & Vitanyi, P. M. B. (2015). Normalized Compression Distance of Multisets with Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(8), 1602–1614. https://doi.org/10.1109/TPAMI.2014.2375175

Colgate, H., S. A., &. Ziock. (2011). *A definition of information, the arrow of information, and its relationship to life. Complexity, 16 (5), 54–62*.

Collier, J. (1990). *Intrinsic information. Information, Language and Cognition: Vancouver Studies in Cognitive Science, 1, 390–409*.

Colombo, M. (2017). Experimental Philosophy of Explanation Rising: The Case for a Plurality of Concepts of *Explanation*. *Cognitive Science*, *41*(2), 503–517. https://doi.org/10.1111/cogs.12340

Coltheart, M. (2012). The cognitive level of explanation. *Australian Journal of Psychology*, *64*(1), 11–18. https://doi.org/10.1111/j.1742-9536.2011.00045.x

Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier?. In *ECAI 2012* (pp. 21-26). IOS Press.

Cooper, R. P., & Peebles, D. (2015). Beyond Single-Level Accounts: The Role of Cognitive Architectures in Cognitive Scientific Explanation. *Topics in Cognitive Science*, *7*(2), 243–258. https://doi.org/10.1111/tops.12132

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica*, *138*(2), 289–301. https://doi.org/10.1016/j.actpsy.2011.08.002

Cover, J. A., T. M., &. Thomas. (2012). *Elements of information theory. John Wiley & Sons*.

Craver, C. F. (2007a). *Constitutive Explanatory Relevance. Journal of Philosophical Research, 32, 3–20. doi: 10.5840/jpr20073241*.

Craver, C. F. (2007b). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Oxford: Oxford University Press*.

Cummins, N., R., Putnam, H., &. Block. (1996). *Representations, targets, and attitudes. MIT press*.

Dale, R. (2008). The possibility of a pluralist cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 155–179. https://doi.org/10.1080/09528130802319078

Dale, R., Dietrich, E., & Chemero, A. (2009). Explanatory Pluralism in Cognitive Science. *Cognitive Science*, *33*(5), 739–742. https://doi.org/10.1111/j.1551-6709.2009.01042.x

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience. Computational and Mathematical Modeling of Neural Systems*. MIT Press.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155–185. https://doi.org/10.1007/s11229-013-0247-7

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). *Is neuroimaging measuring information in the brain?* https://doi.org/10.3758/s13423-016-1002-0

Decock, I., L., &. Douven. (2010). *Similarity After Goodman. Review of Philosophy and Psychology, 2 (1), 61–75. doi: 10.1007/s13164-010-0035-y*.

Dennett, D. C. (1991). *Real Patterns. Journal of Philosophy, 88 (1), 27–51.*

Dijkstra, E.W. (1984) *The Threats to Computing Science (EWD898)*, E.W. Dijkstra Archive. USA: Center for American History, University of Texas at Austin,

http://www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD898.html, *archived at* https://perma.cc/ZU8Y-26TY.

Draaisma, D. (2000). *Metaphors of Memory: A History of Ideas about the Mind*. Cambridge University Press.

Drayson, Z. (2014). The Personal/Subpersonal Distinction. *Philosophy Compass*, *9*(5), 338–346. https://doi.org/10.1111/phc3.12124

Dretske, F. (1982). *The informational character of representations. Behavioral and Brain Sciences, 5 (3), 376–377.*

Dunn, J. M. (2008). Information in computer science. W P. Adriaans & J. van Benthem (Eds.) (Red.), *Philosophy of Information: T. Volume 8* (s. 581–608). Elsevier.

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055–3068.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*(1), 115–135. https://doi.org/10/gdvhwj

Egan, F. (2020). A Deflationary Account of Mental Representation. In J. Smortchkova, K. Dołęga, & T. Schlicht (Eds.), *What Are Mental Representations?* (pp. 26–53). Oxford University Press.

Eliasmith, C. (1996). The third contender: A critical examination of the Dynamicist theory of cognition. *Philosophical Psychology*, *9*(4), 441–463. https://doi.org/10.1080/09515089608573194

Epstein, R. (2016, maja 18). *The empty brain*. Aeon. https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Everson, S. (1997). *Aristotle on perception*. Oxford University Press.

Facchin, M. (2021). Structural representations do not meet the job description challenge. *Synthese, 199 (3–4), 5479–5508. doi: 10.1007/s11229-021-03032-8*.

Facchin, M. (2023). Neural representations unobserved—Or: A dilemma for the cognitive neuroscience revolution. *Synthese, 203 (1), 7. doi: 10.1007/s11229-023-04418-6*.

Facchin, M. (2024). Maps, Simulations, Spaces and Dynamics: On Distinguishing Types of Structural Representations. *Erkenntnis. doi: 10.1007/s10670-024-00831-6*.

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1165622

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Josa a, 4 (12), 2379–2394*.

Firestone, C. (2020). Performance vs. Competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Flack, J. C. (2017). Coarse-graining as a downward causation mechanism. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *375*(2109), 20160338. https://doi.org/10.1098/rsta.2016.0338

Flor, N. V., & Hutchins, E. L. (1991). A case study of team programming during perfective software maintenance. W J. Koenemann-Belliveau, T. G. Moher, & S. P. Robertson (Red.), *Empirical studies of programmers: Fourth workshop* (s. 36–59). Ablex Publishing Corporation.

Floridi, L. (2013). *The philosophy of information*. Oxford University Press.

Fodor, J. A. (1975). *The language of thought*. Crowell.

Fodor, J. A. (1984). Semantics, Wisconsin style. *Synthese*, *59*(3), 231–250. https://doi.org/10.1007/BF00869335

Fodor, J. A. (1986). Why Paramecia Don't Have Mental Representations. *Midwest Studies In Philosophy*, *10*(1), 3–23. https://doi.org/10.1111/j.1475-4975.1987.tb00532.x

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Fodor, J. A. (1990a). *A theory of content and other essays. The MIT press*.

Fodor, J. A. (1990b). Information and representation. W P. Hanson (Red.), *Information, Language, and Cognition* (s. 175–190). University of British Columbia Press.

Fodor, J. A. (1999). *Jerry Fodor · Diary: Why the brain? · LRB 30 September 1999. London Review of Books, 21 (19). Https://www.lrb.co.uk/the-paper/v21/n19/jerry-fodor/diary*.

Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception?: Some reflections on Gibson's "ecological approach". *Cognition*, *9*(2), 139–196. https://doi.org/10.1016/0010-0277(81)90009-3

Földiák, M. P., P., &. Young. (1998). *Sparse coding in the primate cortex, The handbook of brain theory and neural networks.* MIT Press, Cambridge, MA.

Franks, B. (1995). On Explanation in the Cognitive Sciences: Competence, Idealization, and the Failure of the Classical Cascade. *The British Journal for the Philosophy of Science*, *46*(4), 475–502. https://doi.org/10.1093/bjps/46.4.475

Fresco, M., N., &. Miłkowski. (2021). Mechanistic Computational Individuation without Biting the Bullet. *The British Journal for the Philosophy of Science, 72 (2), 431–438. doi: 10.1093/bjps/axz005*.

Fresco, N. (2014). *Physical computation and cognitive science. Springer*.

Fresco, N., Copeland, B. J., & Wolf, M. J. (2021). The indeterminacy of computation. *Synthese*, *199*(5), 12753–12775. https://doi.org/10.1007/s11229-021-03352-9

Friston, J., K., Mattout, J., &. Kilner. (2011). Action understanding and active inference. *Biological Cybernetics, 104 (1–2), 137–160*.

Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*(4), 1273–1302.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind.* Oxford University Press.

Gallistel, C. R. (2019). Our understanding of neural codes rests on Shannon's foundations. *Behavioral and Brain Sciences*, *42*, e226. https://doi.org/10.1017/S0140525X19001249

Gallistel, C. R., & King, A. P. (2011). *Memory and the computational brain: Why cognitive science will transform neuroscience*. John Wiley & Sons.

Galton, F. (1879). Composite portraits, made by combining those of many different persons into a single resultant figure. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *8*, 132–144.

Gärdenfors, P. (1993). The emergence of meaning. *Linguistics and Philosophy*, *16*(3), 285–309. https://doi.org/10.1007/BF00985971

Gärdenfors, P. (2000). *Conceptual spaces* (Vol. 3). Cambridge, MA: MIT press.

Gärdenfors, P. (2005). The detachment of thought. *The mind as a scientific subject: between brain and culture*, 323–341.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724. https://doi.org/10.1016/S0042-6989(00)00277-7

Gent, T. C., Bandarabadi, M., Herrera, C. G., & Adamantidis, A. R. (2018). Thalamic dual control of sleep and wakefulness. *Nature Neuroscience*, *21*(7), 974–984. https://doi.org/10.1038/s41593-018-0164-7

Gibson, J. J. (1950). *The perception of the visual world.* Houghton Mifflin.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Psychology Press.

Gładziejewski, M., P., &. Miłkowski. (2017). *Structural representations: Causally relevant and different from detectors. Biology & Philosophy, 32 (3), 337–355.*

Gładziejewski, P. (2015a). *Explaining Cognitive Phenomena with Internal Representations: A Mechanistic Perspective. Studies in Logic, Grammar and Rhetoric, 40 (1), 63–90. doi: 10.1515/slgr-2015-0004.*

Gładziejewski, P. (2015b). *Wyjaśnianie za pomocą reprezentacji mentalnych: Perspektywa mechanistyczna.* Warszawa: Fundacja na rzecz Nauki Polskiej.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese,* 193 (2), 559–582.

Godfrey-Smith, P. (2011). Signals: Evolution, Learning, and Information, by Brian Skyrms.

Godfrey-Smith, P., & Sterelny, K. (2007). Biological Information. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2007). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2007/entries/information-biological/

Gold, I., & Stoljar, D. (1999). A neuron doctrine in the philosophy of neuroscience. *Behavioral and Brain Sciences*, *22*(5), 809–830. https://doi.org/10.1017/S0140525X99002198

Golonka, S., & Wilson, A. D. (2012). Gibson's ecological approach-a model for the benefits of a theory driven psychology. *Avant*, *3*(2), 40–53.

Goodfellow, A., I., Bengio, Y., &. Courville. (2016). *Deep learning. MIT press*.

Goodman, N. (1972). *Problems and projects.* Bobbs-Merrill.

Google Developers. (2018). *Machine Learning Crash Course with TensorFlow APIs*. https://developers.google.com/machine-learning/crash-course

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438.

Gregory, R. L. (1966). *Eye and brain: The psychology of seeing*. McGraw-Hill.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388. https://doi.org/10.2307/2182440

Griffiths, P. E. (2001). Genetic Information: A Metaphor in Search of a Theory. *Philosophy of Science*, *68*(3), 394–412. https://doi.org/10.1086/392891

Grünwald, A. (2007). *The minimum description length principle.* MIT press.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge University Press.

Hacohen, O. (2022). The problem with appealing to history in defining neural representations. *European Journal for Philosophy of Science, 12 (3), 45. doi: 10.1007/s13194-022-00473-x*.

Haley, P. J., & Soloway, D. (1992, January 1). *Extrapolation limitations of multilayer feedforward neural networks*. https://ntrs.nasa.gov/citations/19930053019

Harnad, S. (2019). Codes, communication and cognition. *Behavioral and Brain Sciences*, *42*, e231. https://doi.org/10.1017/S0140525X19001481

Hatfield, G. (2002). *Perception as unconscious inference.* Wiley Online Library.

Haugeland, J. (2019). Representational genera. In: *Philosophy and the computer* (p. 105–134). Routledge.

Haxby, S. I., J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., ... Rapoport. (1991). Dissociation of object and spatial visual processing pathways

in human extrastriate cortex. *Proceedings of the National Academy of Sciences,* 88 (5), 1621–1625. doi: 10.1073/pnas.88.5.1621.

Helmholtz, H. von. (1867). *Treatise on physiological optics*, vol. III, trans. And ed. JPC Southall. Dover. (Translated from the 3rd German edition, English edition 1962).

Hintikka, J. (1970). *Information, Deduction, and the A Priori*. https://doi.org/10.2307/2214318

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2018, September 27). *Learning deep representations by mutual information estimation and maximization*. International Conference on Learning Representations. https://openreview.net/forum?id=Bklr3jocKX

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840–6851.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hoffman, D. D. (2016). The Interface Theory of Perception. *Current Directions in Psychological Science*, *25*(3), 157–161. https://doi.org/10.1177/0963721416639702

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 15037. https://doi.org/10.1038/ncomms15037

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*, 106–154.

Huttegger, K. J., S. M., Skyrms, B., Smead, R. &. Zollman. (2010). Evolutionary dynamics of Lewis signaling games: Signaling systems vs. Partial pooling. *Synthese, 172 (1), 177*.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content. MIT Press*.

Hyvärinen, P. O., A., Hurri, J., &. Hoyer. (2009). *Natural image statistics: A probabilistic approach to early computational vision.* (Vol. 39). Springer Science & Business Media.

Isaac, A. M. C. (2019). The Semantics Latent in Shannon Information. *The British Journal for the Philosophy of Science*, 70 (1), 103–125. DOI: 10.1093/bjps/axx029.

James, W. (1890). The principles of psychology. *Henry Holt*.

Jozwik, M., K. M., Kriegeskorte, N., &. Mur. (2016). *Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia, 83, 201–226. DOI: 10.1016/j.neuropsychologia.2015.10.023*.

Kalkman, D. (2019). New problems for defining animal communication in informational terms. *Synthese*, *196*(8), 3319–3336. https://doi.org/10.1007/s11229-017-1598-2

Kaminski, M. J., & Blinowska, K. J. (1991). A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, *65*(3), 203–210. https://doi.org/10.1007/BF00198091

Kanizsa, G. (1985). Seeing and thinking. *Acta Psychologica, 59 (1), 23–33*.

Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, *78*(4), 601–627. https://doi.org/10.1086/661755

Kelso, J. A. S., & Engstrøm, D. A. (2008). *The complementary nature*. MIT.

Kelso, J. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT press.

Keren, G., & Schul, Y. (2009). Two Is Not Always Better Than One: A Critical Evaluation of Two-System Theories. *Perspectives on Psychological Science*, *4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x

Kiani, K., R., Esteky, H., Mirpour, K., &. Tanaka. (2007). Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex. *Journal of Neurophysiology, 97 (6), 4296–4309. doi: 10.1152/jn.00024.2007*.

Kiefer, A. B. (2017). Literal Perceptual Inference. *Philosophy and Predictive Processing*. https://doi.org/10.15502/9783958573185

Kim, N.-G., Turvey, M. T., & Carello, C. (1993). Optical information about the severity of upcoming contacts. *Journal of Experimental Psychology: Human perception and performance*, *19*(1), 179.

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. ICLR 2014. https://openreview.net/forum?id=33X9fd2-9FyZd

Kinsey, S. (2022). *A Defense of Algorithmic Homuncularism*. https://philpapers.org/rec/KINADO

Kirsh, D. (1992). When is Information Explicitly Represented? *The Vancouver Studies in Cognitive Science*, 340–365.

Kiverstein, J., & Rietveld, E. (2018). Reconceiving representation-hungry cognition: An ecological-enactive proposal. *Adaptive Behavior*, *26*(4), 147–163. https://doi.org/10.1177/1059712318772778

Klein, B., & Hoel, E. (2020). The Emergence of Informative Higher Scales in Complex Networks. *Complexity*, *2020*(1), 8932526. https://doi.org/10.1155/2020/8932526

Kohár, M. (2023). *Neural machines: A defense of non-representationalism in cognitive neuroscience. Cham: Springer*.

Korbak, T. (2015). Scaffolded minds and the evolution of content in signaling pathways. *Studies in Logic, Grammar and Rhetoric, 41 (1), 89–103*.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in cognitive sciences*, *17*(8), 401–412.

Kriegeskorte, P. A., N., Mur, M., &. Bandettini. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. doi: 10.3389/neuro.06.004.2008.

Kriegeskorte, P. K., N., &. Douglas. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21 (9), 1148–1160.

Kriegeskorte, P., N., &. Bandettini. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, 38 (4), 649–662. doi: 10.1016/j.neuroimage.2007.02.022.

Kriegeskorte, P., N., Goebel, R., &. Bandettini. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103 (10), 3863–3868. DOI: 10.1073/pnas.0600244103.

Krstić, V. (2025). We Should Move on from Signalling-Based Analyses of Biological Deception. *Erkenntnis*, *90*(2), 545–565. https://doi.org/10.1007/s10670-023-00719-x

Kuffler, S. W. (1953). Discharge Patterns And Functional Organization Of Mammalian Retina. *Journal of Neurophysiology*, *16*(1), 37–68. https://doi.org/10.1152/jn.1953.16.1.37

Labarrera, C., Deitcher, Y., Dudai, A., Weiner, B., Kaduri Amichai, A., Zylbermann, N., & London, M. (2018). Adrenergic modulation regulates the dendritic excitability of layer 5 pyramidal neurons in vivo. *Cell Reports*, 23, 1034–1044

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965. Vol. 4 Vol. 4* (pp. 91–195). Cambridge University Press.

Laudan, L. (1977). *Progress and Its Problem: Towards a Theory of Scientific Growth*. University of California Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Lande, K. (2019, April 11). Your brain probably is a computer, whatever that means. *Aeon*. https://aeon.co/essays/your-brain-probably-is-a-computer-whatever-that-means

Larkum, M. E., Wu, J., Duverdin, S. A., & Gidon, A. (2022). The Guide to Dendritic Spikes of the Mammalian Cortex *In Vitro* and *In Vivo*. *Neuroscience*, *489*, 15–33. https://doi.org/10.1016/j.neuroscience.2022.02.009

Le, A. Y., Q. V., Karpenko, A., Ngiam, J., &. Ng. (2011). ICA with reconstruction cost for efficient overcomplete feature learning. *Advances in Neural Information Processing Systems, 1017–1025*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, A. Y., Myers, J., & Rabin, G. O. (2022). The structure of analog representation. *Noûs*, *57*(1), 209–237. https://doi.org/10.1111/nous.12404

Lennie, J. A., P., &. Movshon. (2005). Coding of color and form in the geniculostriate visual pathway (invited review). *JOSA A, 22 (10), 2013–2033*.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (2007). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, *47*(11), 1940–1951.

Lewis, D. (2008). *Convention: A philosophical study.* John Wiley & Sons.

Levoy, Marc (director). (2016, August 28). *Marc Levoy—Lectures on Digital Photography—Lecture 1 (21mar16).mp4* [Video recording]. https://www.youtube.com/watch?v=y7HrM-fk_Rc

Li, M., & Vitanyi, P. (1993). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag.

Li, M.-H., X., Lu, H., Zhang, L., Ruan, X., &. Yang. (2013). Saliency detection via dense and sparse reconstruction. *Proceedings of the IEEE International Conference on Computer Vision, 2976–2983*.

Lombardi, O., Holik, F., & Vanni, L. (2016). What is Shannon information? *Synthese*, *193*(7), 1983–2012. https://doi.org/10.1007/s11229-015-0824-z

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Love, B. C. (2021). Levels of biological plausibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1815), 20190632. https://doi.org/10.1098/rstb.2019.0632

Luna, P. (2011). Controlling machines with just the power of thought. *The Lancet Neurology*, *10*(9), 780–781. https://doi.org/10.1016/S1474-4422(11)70180-6

Lupyan, G. (2015). Cognitive Penetrability of Perception in the Age of Prediction: Predictive Systems are Penetrable Systems. *Review Of Philosophy And Psychology*, 6 (4), 547–569.

Lyon, P. (2015). The cognitive cell: Bacterial behavior reconsidered. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00264

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

MacKay, D. J., D. J., &. Mac Kay. (2003). *Information theory, inference and learning algorithms. Cambridge university press*.

Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies*, *155*(1), 117–131. https://doi.org/10.1007/s11098-010-9562-8

Maley, C. J. (2018a). Continuous Neural Spikes and Information Theory. *Review of Philosophy and Psychology*, 1–21. https://doi.org/10/gdtm6m

Maley, C. J. (2018b). Toward Analog Neural Computation. *Minds and Machines*, *28*(1), 77–91. https://doi.org/10.1007/s11023-017-9442-5

Maley, C. J. (2023). Analogue Computation and Representation. *The British Journal for the Philosophy of Science*, *74*(3), 739–769. https://doi.org/10.1086/715031

Maley, C. J. (2024). Computation for cognitive science: Analog versus digital. *WIREs Cognitive Science*, *15*(4), e1679. https://doi.org/10.1002/wcs.1679

Mann, S. F. (2018). Attribution of Information in Animal Interaction. *Biological Theory*, *13*(3), 164–179. https://doi.org/10.1007/s13752-018-0299-5

Mann, S. F. (2020). Consequences of a Functional Account of Information. *Review of Philosophy and Psychology, 11 (3), 669–687. doi: 10.1007/s13164-018-0413-4*.

Mante, W. T., V., Sussillo, D., Shenoy, K. V., &. Newsome. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature, 503 (7474),* 78–84. doi: 10.1038/nature12742.

Marcus, G. (2018). *Deep Learning: A Critical Appraisal*. https://doi.org/10.48550/arXiv.1801.00631

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of*

London. B. Biological Sciences, *200*(1140), 269–294. https://doi.org/10.1098/rspb.1978.0020

Marr, D. (1982). *Vision*. W. H. Freeman and Company.

Marraffa, M., & Paternoster, A. (2013). Functions, levels, and mechanisms: Explanation in cognitive science and its problems. *Theory & Psychology*, *23*(1), 22–45. https://doi.org/10.1177/0959354312451958

Martin, C. B., & Deutscher, M. (1966). Remembering. *The Philosophical Review*, *75*(2), 161–196.

Martineau, K. (2021, lutego 9). *What is generative AI?* IBM Research. https://research.ibm.com/blog/what-is-generative-AI

Martínez, M. (2019a). Representations are rate-distortion sweet spots. *Philosophy of Science*, *86*(5), 1214–1226.

Martínez, M. (2019b). Usefulness drives representations to truth: A family of counterexamples to Hoffman's interface theory of perception. *grazer philosophische studien*, *96*(3), 319–341.

Martínez, M. (2025). The Information-Processing Perspective on Representation. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.11321

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological review*, *97*(2), 225.

Matsumoto, H., M., &. Komatsu. (2005). *Neural responses in the macaque V1 to bar stimuli with various lengths presented on the blind spot. Journal of Physiology, 93 (5), 2374–2387.*

Maturana, H. R., & Varela, F. J. (2012). *Autopoiesis and cognition: The realization of the living* (Vol. 42). Springer Science & Business Media.

Mavlankar, B., A., Chang, C. L., &. Giros. (2006). *Intraframe Prediction with Intraframe Update Step for Motion-Compensated Lifted Wavelet Video Coding. Picture Coding Symposium*.

McHugh, C., & Way, J. (2018). *What is Good Reasoning?*

Meet Pixel 8 and Pixel 8 Pro, our newest phones. (October 4, 2023). Google. https://blog.google/products/pixel/google-pixel-8-pro/

Mendelovici, A. (2013). Reliable misrepresentation and tracking theories of mental representation. *Philosophical Studies*, *165*(2), 421–443. https://doi.org/10.1007/s11098-012-9966-8

Michaelian, K. (2011). Generative memory. *Philosophical Psychology*, *24*(3), 323–342. https://doi.org/10.1080/09515089.2011.559623

Michaelian, K. (2024). *Mental time travel: Episodic memory and our knowledge of the personal past*. MIT Press.

Millhouse, T. (2022). Really Real Patterns. *Australasian Journal of Philosophy*, *100*(4), 664–678. https://doi.org/10.1080/00048402.2021.1941153

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism.* Cambridge, Mass.: The MIT Press.

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy*, 86 (6), 281. doi: 10.2307/2027123.

Millikan, R. G. (1995). Pushmi-Pullyu Representations. *Philosophical Perspectives*, *9*, 185. https://doi.org/10.2307/2214217

Milner, M. A., A. D., &. Goodale. (1995). *The visual brain in action.* Oxford: Oxford University Press.

Miłkowski, M. (2013). *Explaining the Computational Mind.* Cambridge, Mass.: MIT Press.

Miłkowski, M. (2015). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy*, 30 (5), 709–728. doi: 10.1007/s10539-015-9481-3.

Miłkowski, M. (2016a). Function and causal relevance of content. *New Ideas in Psychology*, 40, 94–102. doi: 10.1016/j.newideapsych.2014.12.003.

Miłkowski, M. (2016b). Unification Strategies in Cognitive Science. *Studies in Logic, Grammar and Rhetoric*, *48*(1), 13–33. https://doi.org/10.1515/slgr-2016-0053

Miłkowski, M. (2017a). Szaleństwo, a nie metoda. Uwagi o książce Pawła Gładziejewskiego “Wyjaśnianie za pomocą reprezentacji mentalnych.” *Filozofia Nauki*, 25 (3 (99)), 57–67.

Miłkowski, M. (2017b). The False Dichotomy between Causal Realization and Semantic Computation. *Hybris*, (38), 1–21.

Miłkowski, M. (2023). Correspondence Theory of Semantic Information. *The British Journal for the Philosophy of Science*, 74 (2), 485–510. doi: 10.1086/714804.

Miłkowski, M., & Hohol, M. (2021). Explanations in cognitive science: Unification versus pluralism. *Synthese*, *199*(S1), 1–17. https://doi.org/10.1007/s11229-020-02777-y

Mingard, C., Valle-Pérez, G., Skalse, J., & Louis, A. A. (2021). Is SGD a Bayesian sampler? Well, almost. *Journal of Machine Learning Research*, *22*(79), 1–64.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Montgomery, R. (1995). Explanation and Evaluation in Cognitive Science. *Philosophy of Science*, *62*(2), 261–282. https://doi.org/10.1086/289856

Morgan, A. (2020). Against neuroclassicism: On the perils of armchair neuroscience. *Mind & Language*, https://doi.org/10.1111/mila.12304

Mudrik, L., Faivre, N., & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, *18*(9), 488–496. https://doi.org/10.1016/j.tics.2014.04.009

Mur, N., M., Bandettini, P. A., &. Kriegeskorte. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience, 4 (1), 101–109. doi: 10.1093/scan/nsn044*.

Mur, N., M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., &. Kriegeskorte. (2013). Human Object-Similarity Judgments Reflect and Transcend the Primate-IT Object Representation. *Frontiers in Psychology*, 4. doi: 10.3389/fpsyg.2013.00128.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). *Progress measures for grokking via mechanistic interpretability* (Nr arXiv:2301.05217). arXiv. https://doi.org/10.48550/arXiv.2301.05217

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In: W. G. Chase (ed.), *Visual information processing* (s. 283–308). Academic Press.

Nizami, L. (2019). Information theory is abused in neuroscience. *Cybernetics & Human Knowing*, *26*(4), 47–97.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological review*, *110*(4), 611.

Nortmann, N., Rekauzke, S., Onat, S., König, P., & Jancke, D. (2015). Primary Visual Cortex Represents the Difference Between Past and Present. *Cerebral Cortex*, *25*(6), 1427–1440. https://doi.org/10.1093/cercor/bht318

Olah, A., C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., &. Mordvintsev. (2018). *The building blocks of interpretability. Distill, 3 (3), e10*.

Oliva, A., &. Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42 (3), 145–175.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, *155*, 23–36.

Olshausen, D. J., B. A., &. Field. (1996). *Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381 (6583), 607*.

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. Oxford University Press.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44 (2), 327–352.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010. https://doi.org/10.3758/BF03196730

Osman, M. (2018). Persistent Maladies: The Case of Two-Mind Syndrome. *Trends in Cognitive Sciences*, *22*(4), 276–277. https://doi.org/10/gc5tvb

Pais-Vieira, M., Lebedev, M., Kunicki, C., Wang, J., & Nicolelis, M. A. (2013). A brain-to-brain interface for real-time sharing of sensorimotor information. *Scientific reports*, *3*(1), 1319.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Clarendon Press.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* MIT press.

Palmer, A. (2018, January 4). How brains and machines can be made to work together. *The Economist*. https://www.economist.com/technology-quarterly/2018/01/04/how-brains-and-machines-can-be-made-to-work-together

Papayannopoulos, P., Fresco, N., & Shagrir, O. (2022). Computational indeterminacy and explanations in cognitive science. *Biology & Philosophy*, *37*(6), 47. https://doi.org/10.1007/s10539-022-09877-8

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon Emissions and Large Neural Network Training* (Wersja 3). arXiv. https://doi.org/10.48550/ARXIV.2104.10350

Pautz, A. (2019). What is the Integrated Information Theory of Consciousness?*: A Catalogue of Questions. Journal of Consciousness Studies, 26(1-2), 188-215.*

Pavese, C. (2021). Practical concepts and productive reasoning. *Synthese*, *199*(3), 7659-7688

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205. http://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html

Pero, F., & Suárez, M. (2016). Varieties of misrepresentation and homomorphism. *European Journal for Philosophy of Science*, *6*(1), 71–90. https://doi.org/10.1007/s13194-015-0125-x

Pezzulo, G., Parr, T., & Friston, K. (2022). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844). https://royalsocietypublishing.org/rstb/article/377/1844/20200531/108758

Phillips, W. A. (2023). *The cooperative neuron: Cellular foundations of mental life*. Oxford University Press.

Piccinini, C., G., &. Craver. (2011). *Integrating psychology and neuroscience: Functional analyses as mechanism sketches. Synthese, 183 (3), 283–311. doi: 10.1007/s11229-011-9898-4*.

Piccinini, G. (2015). *Physical computation: A mechanistic account.* Oxford University Press.

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

Piccinini, G. (2022). *Situated Neural Representations: Solving the Problems of Content. Frontiers in Neurorobotics, 16, 846979. doi: 10.3389/fnbot.2022.846979*.

Piccinini, G., & Bahar, S. (2013). Neural Computation and the Computational Theory of Cognition: Cognitive Science. *Cognitive Science*, *37*(3), 453–488. https://doi.org/10.1111/cogs.12012

Pierce, J. R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications.

Pitcher, L. G., D., &. Ungerleider. (2021). *Evidence for a Third Visual Pathway Specialized for Social Perception. Trends in Cognitive Sciences, 25 (2), 100–110. doi: 10.1016/j.tics.2020.11.006*.

Plebe, V. M., A., &. De La Cruz. (2016). *Neurosemantics.* Cham: Springer International Publishing.

Poldrack, R. A. (2018). *The new mind readers: What neuroimaging can and cannot reveal about our thoughts*. Princeton University Press.

Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1–2), 1307–1325. https://doi.org/10.1007/s11229-020-02793-y

Potochnik, A. (2010). Levels of Explanation Reconceived. *Philosophy of Science*, *77*(1), 59–72. https://doi.org/10.1086/650208

Potochnik, A., & Sanches De Oliveira, G. (2020). Patterns in Cognitive Phenomena and Pluralism of Explanatory Styles. *Topics in Cognitive Science*, *12*(4), 1306–1320. https://doi.org/10.1111/tops.12481

Proudfoot, D., & Copeland, J. (2019). Turing and the First Electronic Brains: What the Papers Said. In M. Colombo & M. Sprevak (ed.), *Routledge Handbook of the Computational Mind* (s. 23–37). Routledge.

Putnam, H. (1960). Minds and machines. In S. Hook (ed.), *Dimensions of Mind* (s. 148–179). New York University Press.

Pylyshyn, Z. (1999). *Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. Behavioral and Brain Sciences, 22 (3), 341–365.*

Raftopoulos, A. (2020). Object individuation by iconic content: How is numerosity represented in iconic representation? *Rivista Internazionale Di Filosofia e Psicologia*, *11*(1), 42–70. https://doi.org/10.4453/rifp.2020.0003

Ramirez, S., Liu, X., Lin, P.-A., Suh, J., Pignatelli, M., Redondo, R. L., Ryan, T. J., & Tonegawa, S. (2013). Creating a False Memory in the Hippocampus. *Science*, *341*(6144), 387–391. https://doi.org/10.1126/science.1239073

Ramsey, W. M. (2007). *Representation reconsidered. Cambridge University Press.*

Ramsey, W. (2020). Defending Representation Realism. In J. Smortchkova, K. Dołęga, & T. Schlicht (Eds.), *What Are Mental Representations?* (pp. 54–78). Oxford University Press.

Rathkopf, B., C., Heinrichs, J. H., &. Heinrichs. (2023). *Can we read minds by imaging brains? Philosophical Psychology, 36 (2), 221–246. doi: 10.1080/09515089.2022.2041590*.

Rathkopf, C. (2017a). Neural information and the problem of objectivity. *Biology & Philosophy, 32 (3), 321–336*.

Rathkopf, C. (2017b). *What Kind of Information is Brain Information? Topoi, 1–8. https://doi.org/10.1007/s11245-017-9512-6*.

Rączaszek-Leonardi, J., & Scott Kelso, J. A. (2008). Reconciling symbolic and dynamic aspects of language: Toward a dynamic psycholinguistics. *New Ideas in Psychology*, *26*(2), 193–207. https://doi.org/10.1016/j.newideapsych.2007.07.003

Recanati, F. (2012). *Mental Files*. Oxford University Press.

Rescorla, M. (2014). Can perception halt the regress of justifications. *Ad infinitum. New essays on epistemological infinitism*, 179–200.

Rescorla, M. (2015). *The computational theory of mind*. https://plato.stanford.edu/entries/computational-mind/?utm_source=the-augmented-human.beehiiv.com&utm_medium=referral&utm_campaign=rewriting-the-ai-myth

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C.,

… Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Rieke, F., Warland, D., Van Steveninck, R. de R., & Bialek, W. (1999). *Spikes: Exploring the neural code*. MIT press.

Ritchie, C., J. B., Kaplan, D. M., &. Klein. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science, 70 (2), 581–607. doi: 10.1093/bjps/axx023*.

Ritchie, H. P., J. B., Lee Masson, H., Bracci, S., &. Op de Beeck. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *NeuroImage, 245, 118686. doi: 10.1016/j.neuroimage.2021.118686*.

Rock, I. (1983). *The Logic of Perception*. MIT Press.

Rolls, E. T. (2021). *Brain Computations: What and How*. Oxford University Press.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton University Press.

Roskies, A. L. (2021). Representational similarity analysis in neuroimaging: Proxy vehicles and provisional representations. *Synthese.* doi: 10.1007/s11229-021-03052-4.

Ross, L. N., & Bassett, D. S. (2024). Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, *25*(2), 81–90. https://doi.org/10.1038/s41583-023-00778-7

Salis, P. (2023). Discursive pluralism: Inferentialist expressivism and the integration challenge. *Metaphilosophy*, *54*(5), 717–733. https://doi.org/10.1111/meta.12655

Scerri, E. (2024). Laws of nature according to some philosophers of science and according to chemists. *Foundations of Chemistry*, *26*(3), 327–341. https://doi.org/10.1007/s10698-024-09517-x

Schalk, N., G., Kapeller, C., Guger, C., Ogawa, H., Hiroshima, S., Lafer-Sousa, R., Saygin, Z. M., Kamada, K., &. Kanwisher. (2017). Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proceedings of the National Academy of Sciences,* 114 (46), 12285–12290.

Schmidhuber, J. (1990). *Making the world differentiable: On using self supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments* (T. 126). Inst. für Informatik. https://people.idsia.ch/~juergen/FKI-126-90ocr.pdf

Schuman, B., Dellal, S., Prönneke, A., Machold, R., & Rudy, B. (2021). Neocortical Layer 1: An Elegant Solution to Top-Down and Bottom-Up Integration. *Annual Review of Neuroscience*, *44*(Volume 44, 2021), 221–252. https://doi.org/10.1146/annurev-neuro-100520-012117

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54–63. https://doi.org/10.1145/3381831

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(03), 1–19. https://doi.org/10.1017/S0140525X00005756

Sejnowski, T. J. (2018). *The deep learning revolution*. The MIT Press.

Sekuler, S. E., A. B., &. Palmer. (1992). Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General, 121 (1), 95*.

Sellars, W. (2007). *In the space of reasons: Selected essays of Wilfrid Sellars*. Harvard University Press.

Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious Inferences: Reply to Sun and Firestone on the Dark Room Problem. *Trends in Cognitive Sciences*, *24*(9), 681–683. https://doi.org/10.1016/j.tics.2020.05.011

Seung, S. (2012). *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt.

Shagrir, O. (2022). *The nature of physical computation*. Oxford University Press.

Shannon, C. E. (1956). *The bandwagon. IRE Transactions on Information Theory, 2 (1), 3*.

Shapiro, L. A. (1997). *Junk Representations. The British Journal for the Philosophy of Science, 48 (3), 345–361. doi: 10.1093/bjps/48.3.345*.

Shapiro, L., & Spaulding, S. (2025). Embodied Cognition. W E. N. Zalta & U. Nodelman (Red.), *The Stanford Encyclopedia of Philosophy* (Summer 2025). Metaphysics

Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2025/entries/embodied-cognition/

Shea, N. (2007). *Content and Its Vehicles in Connectionist Systems. Mind & Language, 22 (3), 246–269. doi: 10.1111/j.1468-0017.2007.00308.x*.

Shea, N. (2018). *Representation in cognitive science. New York, NY: Oxford University Press*.

Shea, N., Godfrey-Smith, P., & Cao, R. (2018). Content in Simple Signalling Systems. *The British Journal for the Philosophy of Science*, *69*(4), 1009–1035. https://doi.org/10.1093/bjps/axw036

Shramko, Y., & Wansing, H. (2019). The nature of entailment: An informational approach. *Synthese*. https://doi.org/10.1007/s11229-019-02474-5

Skyrms, B. (2010). Signals: Evolution, learning, and information. *Oxford University Press*.

Sloman, A. (1989). On designing a visual system (Towards a Gibsonian computational model of vision). *Journal of Experimental & Theoretical Artificial Intelligence*, *1*(4), 289–337. https://doi.org/10.1080/09528138908953711

Sloman, A. (2010). What's information, for an organism or intelligent machine? How can a machine or organism mean? W G. Dodig-Crnkovic & M. Burgin (Red.), *Information and Computation*. World Scientific Publishing.

Sowa, J. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole.

Sprevak, M., & Colombo, M. (eds.). (2019). *The Routledge handbook of the computational mind*. Routledge.

Stafford, J., Whyatt, C., & Craig, C. M. (2019). Age-related differences in the perception of gap affordances: Impact of standardized action capabilities on road-crossing judgements. *Accident Analysis & Prevention*, *129*, 21–29. https://doi.org/10.1016/j.aap.2019.05.001

Steinbock, A. J. (1995). Generativity and generative phenomenology. *Husserl Studies*, *12*(1), 55–79.

Steinert-Threlkeld, S., & Szymanik, J. (2019). Learnability and semantic universals. *Semantics and Pragmatics*, *12*(4), 1–39. https://doi.org/10.3765/sp.12.4

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. The MIT Press.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Sturz, B. R., Boyer, T. W., Magnotti, J. F., & Bodily, K. D. (2017). Do eye movements during shape discrimination reveal an underlying geometric structure? *Animal Behavior and Cognition*, *4*(3), 267–285. https://doi.org/10.26451/abc.04.03.06.2017

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449–508. https://doi.org/10.1007/BF00499820

Tarski, A. (1933). *Pojęcie prawdy w językach nauk dedukcyjnych*. Towarzystwo Naukowe Warszawskie.

Teghtsoonian, M. (1965). The Judgment of Size. *The American Journal of Psychology*, *78*(3), 392. https://doi.org/10.2307/1420573

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. MIT press.

Thomson, G., E., &. Piccinini. (2018). *Neural Representations Observed. Minds and Machines, 28 (1), 191–235. doi: 10.1007/s11023-018-9459-4*.

Thornton, J. P., M. A., &. Mitchell. (2018). Theories of Person Perception Predict Patterns of Neural Activity During Mentalizing. *Cerebral Cortex*, 28 (10), 3505–3520. doi: 10.1093/cercor/bhx216.

Timme, C., N. M., &. Lapish. (2018). *A tutorial for information theory in neuroscience. ENeuro, 5 (3)*.

Tishby, N., N., &. Zaslavsky. (2015). *Deep learning and the information bottleneck principle. Information Theory Workshop (ITW), 2015 IEEE, 1–5*.

Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, *5*(1). https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2015). Integrated information theory. In *Scholarpedia* (Vol. 10, p. 4164). http://www.scholarpedia.org/article/Integrated_information_theory

Tononi, G., & Sporns, O. (2003). Measuring information integration. *BMC Neuroscience*, *4*(1), 31. https://doi.org/10.1186/1471-2202-4-31

Tootell, S. L., R. B., Switkes, E., Silverman, M. S., &. Hamilton. (1988). *Functional anatomy of macaque striate cortex. II. Retinotopic organization. Journal of Neuroscience, 8 (5), 1531–1568*.

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLOS Computational Biology*, *16*(4), e1007805. https://doi.org/10.1371/journal.pcbi.1007805

Turvey, M. T. (2018). *Lectures on perception: An ecological perspective*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780429443879/lectures-perception-michael-turvey

Turvey, M. T., & Shaw, R. E. (1995). Toward an ecological physics and a physical psychology. *The science of the mind: 2001 and beyond*, 144–169.

Tversky, A. (1977). *Features of Similarity. Psychological Review, 84 (4), 327–352*.

Tye, M. (1991). Representation in Pictorialism and Connectionism. W T. Horgan & J. Tienson (Red.), *Connectionism and the Philosophy of Mind* (T. 9, s. 309–330). Springer Netherlands. http://dx.doi.org/10.1007/978-94-011-3524-5_14

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain.* The MIT press. https://psycnet.apa.org/record/2001-18884-000

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. The MIT Press.

van Dijk, L., Withagen, R., & Bongers, R. M. (2015). Information without content: A Gibsonian reply to enactivists' worries. *Cognition*, *134*, 210–214. https://doi.org/10.1016/j.cognition.2014.10.012

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation?: *Journal of Philosophy*, *92*(7), 345–381. https://doi.org/10.2307/2941061

Van Rooij, I., Wright, C. D., & Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. *Synthese*, *187*(2), 471–487. https://doi.org/10.1007/s11229-010-9847-7

Villalobos, M., & Dewhurst J., (2017). Why post-cognitivism does not (necessarily) entail anti-computationalism. *Adaptive Behavior, 25 (3), 117–128. doi: 10.1177/1059712317710496*.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0. http://openaccess.thecvf.com/content_eccv_2018_workshops/w25/html/Wang_ESRGAN_Enhanced_Super-Resolution_Generative_Adversarial_Networks_ECCVW_2018_paper.html

Weaver, W., & Shannon, C. (1964). *The mathematical theory of communication*. University of Illinois Press.

Wilkenfeld, D. A. (2019). *Understanding as compression. Philosophical Studies, 176 (10), 2807–2831.*

Wilkenfeld, D. A., & Lombrozo, T. (2020). Explanation classification depends on understanding: Extending the epistemic side-effect effect. *Synthese*, *197*(6), 2565–2592. https://doi.org/10.1007/s11229-018-1835-3

Wilkinson, S. (2014). Levels and kinds of explanation: Lessons from neuropsychiatry. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00373

Williams, J. J., & Lombrozo, T. (2010). The Role of Explanation in Discovery and Generalization: Evidence From Category Learning. *Cognitive Science*, *34*(5), 776–806. https://doi.org/10.1111/j.1551-6709.2010.01113.x

Willett, F. R., Deo, D. R., Avansino, D. T., Rezaii, P., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2020). Hand Knob Area of Premotor Cortex Represents the Whole Body in a Compositional Way. *Cell*, *181*(2), 396-409.e26. https://doi.org/10.1016/j.cell.2020.02.043

Wittlinger, M., Wehner, R., & Wolf, H. (2006). The Ant Odometer: Stepping on Stilts and Stumps. *Science*, *312*(5782), 1965–1967. https://doi.org/10.1126/science.1126912

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Pergamon*.

Wolff, J. G. (1993). Computing, cognition and information compression. *AI Communications, 6 (2), 107–127*.

Wong, R. (2021, February 29). *Is Samsung's Galaxy S21 Ultra using AI to fake detailed Moon photos?* Input. https://www.inverse.com/input/reviews/is-samsung-galaxy-s21-ultra-using-ai-to-fake-detailed-moon-photos-investigation-super-resolution-analysis

Woollacott, E. (2018, February 2). How to control a machine using your mind. *BBC News*. https://bbc.com/news/business-42894312

Wright, C. (2014). *Comment on Paul Boghossian,"What is inference"*.

Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.-K., Levoy, M., & Milanfar, P. (2019). Handheld multi-frame super-resolution. *ACM Transactions on Graphics*, *38*(4), 1–18. https://doi.org/10.1145/3306346.3323024

Xu, J., L., Zheng, S. &. Jia. (2013). Unnatural lo sparse representation for natural image deblurring. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1107–1114*.

Yamins, J. J., D. L. &. DiCarlo. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19 (3), 356*.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*(2), 201–219. https://doi.org/10.1007/s11098-008-9324-z

Yousif, S. R. (2022). Redundancy and reducibility in the formats of spatial representations. *Perspectives on Psychological Science*, *17*(6), 1778-1793.

Zadeh, L. A. (1965). *Fuzzy sets. Information and Control, 8 (3), 338–353. doi: 10.1016/S0019-9958(65)90241-X.*

Zahnoun, F. (2025). Information in the Brain: From Metaphor to Truth. *Theoria*, *91*(5), e70021. https://doi.org/10.1111/theo.70021

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34 (2), 265-288*.

Zhang, Y. (2022). Algorithmic photography: A case study of the Huawei Moon Mode controversy. *Media, Culture & Society*, *44*(4), 690–705. https://doi.org/10.1177/01634437211064964

Appendix (LLM Lawsuits):

1. The Artists' Class Action:

Andersen v. Stability AI Ltd., No. 3:23-cv-00201 (N.D. Cal. filed Jan. 13, 2023). https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/

2. The New York Times:

The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023). https://dockets.justia.com/docket/new-york/nysdce/1:2023cv11195/612697

3. The Academics (Basbanes & Gage):

Basbanes v. Microsoft Corp., No. 1:24-cv-00084 (S.D.N.Y. filed Jan. 5, 2024). https://dockets.justia.com/docket/new-york/nysdce/1:2024cv00084/613126

4. The Reddit Users (J.L. Class Action):

J.L. v. Alphabet Inc., No. 3:23-cv-03440 (N.D. Cal. filed July 11, 2023). https://dockets.justia.com/docket/california/candce/3:2023cv03440/415223